THE UNIVERSITY
of EDINBURGH

THE UNIVERSITY
of EDINBURGH

Healthy Ageing and Binding Features in Working Memory: Measurement Issues and Potential Boundary Conditions

Stephen Rhodes

PhD Psychology

The University of Edinburgh

2016

Declaration

I hereby declare that this thesis is of my own composition, and that it contains no material previously submitted for the award of any other degree. The work reported in this thesis has been conducted by myself, except where due acknowledgement is made in the text.

Stephen Rhodes

Wednesday 17th August, 2016

Acknowledgments

I would like to thank Bob Logie, Mario Parra and Nelson Cowan for their support and guidance throughout this work. Their knowledge and insight has been invaluable. I am also indebted to the participants in this research, particularly members of the Edinburgh Psychology volunteer panel, for giving up their time. For the opportunity to come and study at Edinburgh, I am extremely grateful to Bob, Ian Deary, Sharon Abrahams and the Centre for Cognitive Ageing and Cognitive Epidemiology. Thanks also to Ullrich Ecker for sharing his abstract polygons for Experiment 9.

My friends, both in Edinburgh and beyond, have been the best support I could ask for, academically and non-academically. In particular, I have had many valuable discussions with Jason Doherty, who has the uncanny ability to detect flaws in any design, and Angela de Bruin, who is everything an early career researcher should aspire to—I'm glad to have them on my side (most of the time). Last but certainly not least, my family have been a continuous source of encouragement and motivation throughout my life. In particular I would not have been able to complete this thesis without the support of my mother, brother and sisters.

Contents

Abstract

Accurate memory for an object or event requires that multiple diverse features are bound together and retained as an integrated representation. There is overwhelming evidence that healthy ageing is accompanied by an *associative deficit* in that older adults struggle to remember relations between items above any deficit exhibited in remembering the items themselves. However, the effect of age on the ability to bind features *within* novel objects (for example, their colour and shape) and retain correct conjunctions over brief intervals is less clear. The relatively small body of work that exists on this topic to-date has suggested no additional working memory impairment for conjunctions of features beyond a general age-related impairment in the ability to temporarily retain features. This is in stark contrast to the feature binding deficit observed in the early stages of Alzheimer's disease. Nevertheless, there have been reports of age-related feature binding deficits in working memory under specific circumstances. Thus a major focus of the present work was to assess these potential boundary conditions.

The change detection paradigm was used throughout this work to examine age-differences in visual working memory. Despite the popularity of this task important issues regarding the way in which working memory is probed have been left unaddressed. Chapter 2 reports three experiments with younger adults comparing two methods of testing recognition memory for features or conjunctions. Contrary to an influential study in the field, it appears that processing multiple items at test does not differentially impact on participants' ability to detect binding changes.

Chapters 3, 4, and 5 report a series of experiments motivated by previous findings of specific age-related feature binding deficits. These experiments, improving on pre-

vious methodology where possible, demonstrate that increasing the amount of time for which items can be studied (Chapter 3) or mixing feature-conjunction changes in trial-blocks with more salient changes to individual features (Chapters 4 and 5) *does not* differentially impact on healthy older adults' ability to detect binding changes. Rather, the argument is made that specific procedural aspects of previous work led to the appearance of deficits that do not generalise. Chapter 5 also addresses the suggestion that healthy ageing specifically affects the retention of item-location conjunctions. The existing evidence for this claim is reviewed, and found wanting, and new data are presented providing evidence against it.

To follow-up on the absence of a deficit for simple feature conjunctions, Chapter 6 contrasts two theoretically distinct binding mechanisms: one for features intrinsic to an object and another for extrinsic, contextual features. Preliminary evidence is reported that the cost associated with retaining pairings of features is specifically pronounced for older adults when the features are extrinsic to each other.

In an attempt to separate out the contribution of working memory capacity and lapses of attention to age-differences in overall task performance, Chapter 7 reports the results of an exploratory analysis using processing models developed in Chapter 2. Analysis of two data sets from Chapters 4 and 5 demonstrates that lapses of attention make an important contribution to differences in change detection performance.

Chapter 8 returns to the issue of measurement in assessing the evidence for specific age-related deficits. Simulations demonstrate that the choice of outcome measure can greatly affect conclusions regarding age-group by condition interactions, suggesting that some previous findings of such interactions in the literature may have been more apparent than real.

In closing the General Discussion relates the present work to current theory regarding feature binding in visual working memory and to the wider literature on binding deficits in healthy and pathological ageing.

Lay Summary

The ability to retain associations between the basic features of objects (e.g. colour, shape, location) over brief intervals appears to be drastically affected by early Alzheimer's disease. By contrast many studies have found that healthy older adults are able to retain conjunctions of features just as well as the component features individually (e.g. shape alone). Thus it has been suggested that this so called 'feature binding' function of working memory may be able to help better discriminate between healthy and pathological ageing. Nevertheless there have been several reports of potential boundary conditions under which healthy older adults may struggle to maintain combinations of features.

Across several studies directly assessing such boundary conditions, participants were required to retain multiple objects over a brief interval in order to detect changes to the individual features or the feature conjunctions. In contrast to some previous reports we find evidence against a disproportionate effect of healthy ageing on feature binding in visual working memory. In addressing these boundary conditions we come across a number of statistical issues in the assessment of age-differences in task performance. Namely if the measure of performance is not clearly justified, erroneous support for an age-related binding deficit may become commonplace. The findings of this thesis strengthen the suggestion that feature binding in working memory is preserved in healthy old age and shows that researchers should be wary when evaluating this issue.

Chapter 1

Introduction

1.1 Overview

Accurate memory for an object or event requires that different attributes are combined to form a coherent, integrated representation. This requires mechanisms of binding responsible for this integration. It has been suggested that healthy ageing impairs this ability to bind disparate elements together leading to weaker, more fragmented memories that are less useful in supporting the reminiscence of previous events. However, it does not appear to be the case that all mechanisms of binding are affected equally; the ability to form associations between distinct items (e.g. face and name) exhibits pronounced decline across the lifespan, whereas maintaining arbitrary conjunctions of features (e.g. colour and shape) from the same object over brief intervals appears to be relatively insensitive to the effects of age. This apparent age-invariance of feature binding stands in stark contrast to a pronounced feature binding deficit observed in early Alzheimer's disease. Nevertheless, there remain conditions under which reliable age-related feature binding deficits have been observed. The present work aims to investigate some of these potential boundary conditions. In doing so we repeatedly find no evidence for a disproportionate effect of age on the ability to form temporary bindings between object features (colour, shape, and location) in working memory. This strengthens the proposal that the efficacy of feature binding functions may distinguish between healthy and pathological

ageing.

1.2 Introduction

Performance on a wide range of cognitive tasks shows marked decline across the adult lifespan (Glisky, 2007; Verhaeghen & Salthouse, 1997). The pervasive effect of age on cognition has led to accounts that focus on changes to fundamental cognitive primitives as the source of this widespread decline. These accounts include those that interpret poor task performance in terms of reduced speed of processing (Salthouse, 1996) or those that propose that older adults are less able to inhibit irrelevant information (Hasher & Zacks, 1988). However, merely noting decline in memory performance with age—that is, accepting the 'dull hypothesis' (Perfect & Maylor, 2000; Logie, Horne, & Pettit, 2015)—is not sufficient to describe age-related change. A large body of evidence clearly shows that the effect of age is not uniform across different hypothetical memory systems and materials.

In the long-term memory literature it is well established that, when retrieving previously encountered information, older adults seem less able to recall the source of the information and use this contextual detail to guide memory judgements (Kausler & Puckett, 1981; McIntyre & Craik, 1987; see Spencer & Raz, 1995 for a review and meta-analysis). For example, McIntyre and Craik (1987) presented younger and older adults with lists of facts that were either read aloud by an experimenter or presented visually on a screen. After a week delay memory for the facts and their encoding context was assessed. They found that, even when recall of facts was well matched between the two age-groups, older adults exhibited a disproportionate deficit in the recall of contextual features accompanying the factual information. This lack of contextual detail is also evident in the subjective experience of older adults during memory tasks. In cued-recall tasks older adults are less likely to respond on the basis of conscious recollection that an item was previously encountered but appear to rely on less precise feelings of knowing that an item was encountered (Mäntylä, 1993; McCabe, Roediger, McDaniel, & Balota, 2009). Thus, older adults appear to have a specific problem with episodic memory—recollection

of past events along with their spatiotemporal context (Tulving, 1972)—whereas context free, semantic memory is relatively spared (P. A. Allen, Sliwinski, Bowie, & Madden, 2002).

Research on working memory has also found heterogeneous age-effects depending on the stimulus materials used. Namely, age-related effects appear to be larger when participants are required to maintain visuospatial material—such as faces, matrix patterns, or coloured shapes—relative to verbal material—like letters, words, or digits (Jenkins, Myerson, Joerding, & Hale, 2000; W. Johnson, Logie, & Brockmole, 2010; Leonards, Ibanez, & Giannakopoulos, 2002; Myerson, Hale, Rhee, & Jenkins, 1999; Myerson, Emery, White, & Hale, 2003; Tubi & Calev, 1989). For example, Leonards et al. (2002) assessed participants aged between 20 and 69 on an n-back task which required concurrent processing and maintenance of letters or images of faces and doors. Task performance for letters showed very little age-related decline suggesting largely preserved storage and rehearsal of verbal material. Conversely there was a strong age-effect for both the face and door stimuli. This has been corroborated more recently by a large internet study of working memory across the life-span. In a sample of over 95,000 participants, W. Johnson et al. (2010) found that short-term retention of digits was largely spared with increasing age, whereas recall of matrix patterns or coloured shapes in different locations showed pronounced decline from the mid-20s onwards (see also, Brockmole & Logie, 2013; Logie & Maylor, 2009; Maylor & Logie, 2010).

Hence, both in the long-term and working memory literatures, age-effects appear to be exacerbated through the use of multifaceted stimuli that require the integration of multiple attributes for accurate representation. It is therefore not surprising that multiple groups have proposed that older adults have a particular problem in *binding* together attributes into a coherent representation (Bayen, Phelps, & Spaniol, 2000; Chalfonte & Johnson, 1996; Naveh-Benjamin, 2000; Sander, Lindenberger, & Werkle-Bergner, 2012; Shing et al., 2010). Notably, Naveh-Benjamin (2000), in postulating the *associative deficit* hypothesis of memory ageing, made the broad suggestion that ageing impairs the representation of any "two mental codes" (pp.

1170). This suggestion has gained a lot of support in the study of long-term recognition memory but, as we outline further below, may be a little too general to understand recent findings from studies of visual working memory (VWM).

1.3 The Associative Deficit

Naveh-Benjamin (2000) tested his associative deficit hypotheses by presenting younger and older adults with pairs of semantically unrelated words for a recognition test following a 90 second filled delay. There was a small effect of age on the ability to discriminate between previously seen words and brand new lures; a small age-related effect on *item* recognition. By contrast there was a much larger effect of age on the ability to discriminate previously seen word pairs among re-paired lures (i.e. words that had appeared at study but paired with other words); a disproportionate effect on *associative* recognition. This *associative deficit* is ubiquitous and has been found with a variety of stimuli, such as pairs of objects (Naveh-Benjamin, Hussain, Guez, & Bar-On, 2003), pairs of faces (Bastin & Van der Linden, 2006; M. G. Rhodes, Castel, & Jacoby, 2008), faces and names (Naveh-Benjamin, Guez, & Shulman, 2004), and the actions performed by specific actors (Kersten, Earles, Curtayne, & Lane, 2008; Old & Naveh-Benjamin, 2008b). Indeed a meta analysis of 90 experiments has clearly shown that the effect of age is larger for associations than for items (Old & Naveh-Benjamin, 2008a).

However, it is important to note that the associative deficit does not appear to affect all aspects of recognition equally. Studies that separately assess the rate of correctly identifying previously seen pairs of items (referred to, in this literature, as *hit rate*) and the rate of incorrectly responding 'old' to recombined items (*false alarm rate*) consistently find a larger age-related effect in the latter (e.g., Bender, Naveh-Benjamin, & Raz, 2010; Castel & Craik, 2003; M. G. Rhodes et al., 2008). That is, older adults are more likely to falsely recognise recombined lures of items that were not presented together during the study period. This has been attributed to an increased reliance on feelings of familiarity in making recognition judgements with age. The individual items that make up a recombined lure each elicit a feeling

of familiarity, whereas recollection of what items appeared together is needed to reject the conjunction lure. This suggestion is corroborated by the effects of repetition on older adults' recognition performance. Repeating pairs of stimuli multiple times during study improves the detection of old pairings (i.e. increases hit rate) but also disproportionately increases the likelihood that older adults will falsely endorse conjunction lures as previously seen (M. G. Rhodes et al., 2008; Kilb & Naveh-Benjamin, 2011). In the following sections we review several accounts of the associative deficit that have been offered in the literature.

The Role of Attention in the Associative Deficit

One popular suggestion is that the integration of contextual features may be particularly demanding of attention relative to item memory and given that older adults appear to have reduced 'processing resources' (Craik, 2006; Craik & Byrd, 1982) they struggle to associate disparate items as efficiently as younger adults. This has led researchers to question whether they can simulate the associative deficit of older adults in younger groups by reducing the availability of attentional resources. However, much of the evidence from dividing attention in younger adults has suggested that this is not the case (Craik, Luo, & Sakuta, 2010; Naveh-Benjamin et al., 2004; Kilb & Naveh-Benjamin, 2007). For example, Craik et al. (2010) report two experiments in which they used a concurrent digit monitoring task to divide the attention of a group of younger adults during the encoding of semantically unrelated noun-scene pairs. This divided attention group were then compared to a group of younger adults under full attention and a group of healthy older adults. Performing the digit monitoring task—looking for two odd numbers in a random digit sequence—had a dramatic effect on recognition performance. However, this was true for both item and associative recognition, thus providing a poor simulation of older adults' performance, for whom item memory is relatively unimpaired but associative recognition is impoverished. They concluded that older adults exhibit an associative deficit *over and above* any recognition memory problems associated with reduced attentional control.

Nevertheless, there have been some studies that have found a disproportionate effect of divided attention on associative recognition (Castel & Craik, 2003; Troyer & Craik, 2000; Troyer, Winocur, Craik, & Moscovitch, 1999). However, in these studies the effect has manifested itself in reduced correct acceptance of old items *as well as* increased false recognition of associative lures. This is qualitatively different from the pattern of performance exhibited by older adults which is apparent primarily in the tendency to accept brand new pairings of previously seen items as 'old' (a tendency to false associative recognition). Kim and Giovanello (2011) have suggested that previous studies have struggled to accurately recreate the associative deficit in younger adults as they have been looking at a general form of attention. Instead they posit a form of relational attention, important for processing associations between distinct items. To tax this relational form of attention they introduced a dual task that required participants to actively process relations—in this case by comparing the perceived ages of two presented faces—whilst encoding unrelated word pairs. This was compared to a similar dual task condition that did not require processing of relations—locate the male face within the pair presented. The relational processing condition, as expected by the authors, had a disproportionate effect on pair recognition relative to item recognition. Further, the pattern of performance exhibited under this condition was very similar to that of a group of older adults, as both showed a tendency to falsely recognise associative lures. These preliminary findings raise the fascinating prospect of a relational form of attention that is disrupted in healthy ageing.

Decline in a relational form of attention is also in line with demonstrations of the associative deficit over very short retention intervals (T. Chen & Naveh-Benjamin, 2012; Hartman & Warren, 2005). For example, T. Chen and Naveh-Benjamin (2012) used a continuous recognition paradigm in which participants were required to monitor a stream of face-scene pairs. Interspersed with these study events there were tests of item memory (*'was this face or scene present before?'*), and associative memory (*'was this face paired with this scene?'*). By varying the number of study events between these test events participants could be probed on a given pairing after nu-

merous delays, including an immediate test. The associative deficit, and tendency to falsely recognise relational lures, was present at all delays and did not appear to be exacerbated by increasing the number of events interspersed between study and test. Combined with the evidence provided by Kim and Giovanello (2011), these findings suggest that the associative deficit arises during the initial attentional selection and encoding of relations in working memory.

Strategy Use and the Associative Deficit

While it has proven difficult to reproduce the associative deficit in younger adults using divided attention, there may be a role for differences in strategy use across the lifespan. When participants incidentally encode items and associations for an unexpected memory test the associative deficit is greatly reduced (Naveh-Benjamin, 2000; Old & Naveh-Benjamin, 2008a), or even disappears (Naveh-Benjamin et al., 2009), relative to conditions in which encoding is deliberate. This suggests that the younger adults may receive an 'associative boost' when deliberately encoding associations through the use of elaborate, possibly attentionally demanding, strategies. Further the fact that the associative deficit is larger for rejecting new combinations of previously seen items relative to accepting old pairs indicates that older adults do better when the environment reinstates the combination for them but may struggle to adopt effective strategies to reject familiar lures (cf. M. G. Rhodes et al., 2008).

However, the role of a strategy difference across age-groups is not clear. One early study found that providing an associative strategy, of linking word pairs into a memorable sentence, almost eradicated the associative deficit (Naveh-Benjamin, Brav, & Levy, 2007), however a second larger study found more modest gains. Shing, Werkle-Bergner, Li, and Lindenberger (2008) collected data from four age groups (children, teenagers, young adults, and older adults) on an associative recognition task for word pairs. Their participants' performance was assessed before and after instruction on an elaborative visual imagery strategy that involved forming a vivid image using the paired words. Older adults benefited from this strategy but their gain was approximately the same as the younger group. The limited literature on

strategy effects makes conclusions difficult to reach. Nevertheless, it is worth noting that the associative deficit found in the non-strategy (i.e. baseline) condition was much more pronounced in the study of Naveh-Benjamin et al. (2007) relative to Shing et al. (2008). It seems possible that the magnitude of associative deficit will be an important limiting factor on the ability to see strategy gains. This remains an important avenue for further research.

In summary, decades of research has shown that older adults are less likely to remember the source of information or the spatio-temporal contextual details surrounding an event. This has been attributed to a specific associative deficit, with memory for individual items left relatively intact with age. It has been suggested that age-related decline in the amount of attention that can be devoted to processing relations or age differences in the use of relational strategies may underlie this deficit. The support for these accounts is mixed, but it is clear that neither can fully account for the associative deficit exhibited by older adults. This has resulted in the recent proposal of a two component framework for understanding the associative deficit. Shing et al. (2010) distinguish between the contributions of an associative component, which serves to bind aspects of an event, and a strategic component, which serves to organise information at encoding and allow the efficient use of cues at retrieval (see also, Moscovitch, 1992). They propose that the associative element is largely automatic and dependent on the integrity of medial temporal brain regions, such as the hippocampus. The hippocampus has been shown via both neuropsychological studies of amnesic patients and functional neuroimaging to be important in forming relations between disparate items (see, Olsen, Moses, Riggs, & Ryan, 2012, for a review). On the other hand the implementation of strategies requires controlled processing drawing more heavily on the frontal lobes. Given the pronounced effect that healthy ageing has on both the hippocampus and frontal lobes, particularly pre-frontal cortex (Raz & Rodrigue, 2006), both of these components are said to contribute to the associative deficit. This two component account is in line with the finding that dual attention studies cannot fully simulate the effects of healthy ageing in younger adults and with the finding that a residual associative

deficit remains following strategy instruction, although the role of strategy needs further clarification.

1.4 Age and Binding Visual Features in Working Memory

As outlined above, there appears to be a disproportionate effect of age on the ability to maintain complex visual stimuli made up of multiple features (e.g. W. Johnson et al., 2010; Leonards et al., 2002). There is a great deal of physiological and psychophysical evidence that different feature dimensions[1], such as colour and form, are processed in largely separate parallel streams, raising the question of how these features are attributed to the same object and 'bound' together (see Section 1.5 for more detail on the binding problem and theories of how it is solved). Therefore, influenced by the long-term memory literature on the associative deficit described in the previous section, some have suggested that older adults may also struggle to bind the various visual features of objects into unitised object representations and retain these in visual working memory (VWM) (Brockmole, Parra, Della Sala, & Logie, 2008; Cowan, Naveh-Benjamin, Kilb, & Saults, 2006; Sander, Lindenberger, & Werkle-Bergner, 2012).

Binding Items to Location

Mitchell and colleagues (Mitchell, Johnson, Raye, Mather, & D'Esposito, 2000; Mitchell, Johnson, Raye, & D'Esposito, 2000) were the first to assess the effect of age on binding in VWM. In their task participants were presented with a sequence of three familiar, highly nameable, objects in different locations on a visible 3×3 grid. Following an 8.5 second retention interval memory for the locations, objects, or object-location conjunctions was tested by a single probe item. As is seen in the associative long-term memory literature, older adults were more likely than

[1]The terminology used here is borrowed from the literature on visual attention. *Feature dimension* refers to a defining quality that stimuli can vary on (e.g. colour), whereas a *feature value* is a specific point on a feature dimension (e.g. blue).

younger adults to accept repaired lures as previously seen; that is, they were more likely to miss binding changes. On the other hand there was no discernible effect of age on the detection of object or location changes alone. In a further experiment, Mitchell, Johnson, Raye, and D'Esposito (2000) had participants complete this task in an fMRI scanner in order to assess age differences in blood-oxygen-level dependent signal change during the task. Crucially, their younger adult group exhibited activation of the left hippocampus specifically during the maintenance of object-location associations in VWM, but no such specific recruitment could be found in the older adults. The authors proposed that hippocampal dysfunction in the older group led to poorer binding of features in working memory, in line with findings in long-term memory (see, Shing et al., 2010, for a review).

This has led some to suggest that binding to location, as opposed to binding the surface features of objects, is specifically affected by age (Brockmole et al., 2008). However, subsequent work on location binding has been less clear (see Chapter 5 for a detailed critique of the statistical evidence offered by Mitchell et al.). Using the change detection task, Olson et al. (2004) required participants to detect whether or not a probed item had changed location between study and test. Their crucial manipulation was to either leave the unprobed items unchanged or shuffle their locations, thereby changing the relative spatial location of items (see also, Jiang, Olson, & Chun, 2000). Older adults appeared to benefit as much as younger adults from preserved spatial configuration, suggesting that there is little effect of age on the ability to bind the relative locations of objects together (see also, Read, Rogers, & Wilson, 2016).

Cowan et al. (2006) used a similar change detection task in which participants retained a variable number of coloured squares in different locations over a 1 second interval. Recognition was probed by re-presenting the initial array with a single item cued by a surrounding circle. This cued item would either match the colour presented at that location (no-change), contain a brand new colour that was no part of the study array (item change), or contain a colour that was previously in a different location (item-location binding change). They also recruited four

age groups (grade 3, grade 5, younger adults and older adults) allowing them to assess the life-span trajectory of binding in VWM. When item and binding change trials were presented in separate blocks (Experiment 2A)—the same set up used by Mitchell and colleagues—older adults were able to perform the change detection discrimination with the same accuracy for binding trials as for item trials. This was seen in a signal detection theory analysis which showed that, despite lower levels of sensitivity relative to younger adults, there was no disproportionate effect of old age on binding. Curiously, however, when item and binding trials were mixed within a block (Experiment 1A) older adults exhibited a clear binding deficit in terms of sensitivity (d'), as older adults were more likely to miss item-location changes. Cowan et al. suggested that, as the more salient item changes were mixed with the less salient changes to binding, the older adults may have been lured into using familiarity based recognition (*'did I encounter that colour?'*) rather than recollection (*'did that colour appear in that position?'*). They proposed that when the different trial types were blocked this processing mode would not support the detection of any binding changes therefore the older group were forced to change tack and rely on a more recollective form of recognition.

Indeed it is the case that many of the studies that have failed to found evidence of an age-related binding deficit in working memory (Bopp & Verhaeghen, 2009; Brockmole et al., 2008, see below for further examples) have used blocked presentation of feature and binding changes. Therefore, the Cowan et al. findings may point to a potential role of test salience and the use of familiarity based recognition in the emergence of working memory binding deficits (see, T. Chen & Naveh-Benjamin, 2012, for a similar argument). Mixing trial types versus blocking them may also reveal strategy differences at encoding; given that in blocked conditions participants know what kind of test to prepare for they may be able to deploy more elaborative encoding strategies, particularly for feature combinations (Mitchell, Johnson, Raye, Mather, & D'Esposito, 2000). Older adults may be less likely to engage in this kind of strategy use when the type of upcoming test is known (Naveh-Benjamin et al., 2007) and may, therefore, not benefit from blocked trials. Despite these prospects

for a role of mixing versus blocking trials other studies have failed to show such an effect. T. Chen and Naveh-Benjamin (2012) compared memory for face-scene pairs under conditions where item and associative changes were presented in separate blocks or mixed in the same trial block and found that this did not modulate the size of the associative deficit. Thus the role of mixing or blocking in the emergence or exacerbation of age-related binding deficits is unclear and further work is clearly needed to address this. In the present work we conduct a series of experiments assessing the binding of simple visual features (colour, shape, location) and the effect of mixing trials on older adults' change detection performance (Chapters 4 and 5).

Recently, there has also been interest in older adults' short-term recall of object-location conjunctions. Peich, Husain, and Bays (2013) used a delayed reproduction task to assess the ability of participants, aged 19 to 77, to recall both the colour and orientation of a probed item. They found a marked increase in the variability of recall error with age, even at set size one (low-load), suggesting that representations become less precise with age, something which previous change detection experiments have overlooked (see also, Noack, Lovden, & Lindenberger, 2012). Crucially they fit a mixture model to their data allowing them to assess age-differences in three sources of recall error; 1. the precision of representation; 2. random guessing when the probed item is outside of VWM; 3. erroneous recall of an un-probed item, also called 'mis-binding' (see, Bays, Catalao, & Husain, 2009; Bays, Wu, & Husain, 2011). Peich et al. found that when three items were presented (high-load) this third component, the probability of recalling a feature from an un-probed location, greatly increased with age, suggesting that older adults were more likely to mis-bind features to location in VWM.

However, it is worth noting that the nature of Peich et al's task was such that when recalling features participants were started off at a random feature value within the circular space and then had to cycle through alternatives until the remembered value appeared. This set-up of the task, as opposed to the version in which participants are presented with all options at once, on a colour wheel for example (W. Zhang & Luck, 2008), could conceivably cause an appreciable amount of out-

put interference. Participants cycling to a previously remembered feature may cycle past a feature from another location and, given the feeling of familiarity, erroneously stop cycling. Older adults appear to be more likely to respond on the basis of feelings of familiarity (e.g. M. G. Rhodes et al., 2008), and thus may be more susceptible to this output interference.

Indeed it seems that, when this output interference factor is controlled for, older adults do not exhibit this tendency to mis-bind. Pertzov, Heider, Liang, and Husain (2015) used a task in which participants were presented with either one or three difficult-to-name fractals. Following a delay of either one or four seconds they were asked which of two items were in the initial array and then were required to relocate the chosen item, using a touch-screen, to its remembered location. In line with Peich et al., they found that the probability of dragging a correctly identified fractal to a location that was previously occupied by another item increased with age. However, they point out that it is possible that participants occasionally guessed which of the two fractals were previously seen before relocating. Given that older adults were less likely to identify previously seen fractals this increase in swap rate could really reflect an increased rate of guessing. Correcting for age differences in correct identification of previously seen items, Pertzov et al. found no evidence that location binding errors increased with age.

In summary, while early work suggested that older adults find remembering *what was where* more difficult than either of these attributes individually, subsequent work has shown that this is not always the case. Much of this discrepancy may come down to the use of highly nameable items (clip-art-like images on a grid) versus simple features where verbal strategies may play less of a role. This is complicated by the fact that Cowan et al. (2006) found a binding deficit for briefly presented pairings of colour and location under conditions where feature and binding trials were mixed. The contrast of mixed versus blocked trials may reveal age-differences in the use of relational encoding strategies or the use of recollection at test. However, other investigations into the role of mixing trials have found no difference relative to blocked trials (T. Chen & Naveh-Benjamin, 2012). Further, it is often suggested

that location, relative to other feature dimensions, holds privileged status in visual attention and possibly in VWM. In Section 1.5 below, we discuss the evidence for a special status for location in VWM and potential reasons to expect a specific effect of healthy ageing on the ability to bind objects to locations.

Binding Surface Features

While the evidence concerning the effect of age on object-location binding has been mixed, somewhat more consistent results have come from studies assessing the binding of surface features, such as shape and colour. Brockmole et al. (2008) used a change detection task in which younger and older participants were required to remember a briefly presented array of coloured shapes over a short blank interval (1 second) in order to detect a change in a subsequent test array. In some blocks of trials participants were required to detect changes to the individual features (either colour only or shape only), and in other blocks they were required to detect changes to the combination of features between the two arrays. Overall, relative to the younger group, change detection performance was poorer in the older group, reflecting reduced VWM capacity with age (see also, Sander, Werkle-Bergner, & Lindenberger, 2011a; Jost, Bryck, Vogel, & Mayr, 2011). However, the older group's performance in the binding condition was not significantly different from the shape only condition, suggesting that older adults are still able to bind features in VWM, with performance limited by the most difficult feature dimension (see also Brockmole & Logie, 2013; Parra, Abrahams, Logie, & Della Sala, 2009). Brockmole et al. (Experiment 3) also provide converging evidence from a recall task that older adults were just as proficient in recalling combinations of features as they were in recalling individual features (see also, van Geldorp, Parra, & Kessels, 2015).

Further studies using the change detection paradigm have largely corroborated Brockmole et al's findings (Brown, Niven, Logie, Rhodes, & Allen, 2016; Isella, Molteni, Mapelli, & Ferrarese, 2015; Read et al., 2016). For example, Isella et al. (2015) used a similar procedure to Brockmole et al. in larger groups of younger and older adults and also found no evidence of an age by condition interaction

in proportion correct. Further, Brown et al. (2016) conducted a series of change detection experiments comparing younger and older adults' binding performance under various conditions. In two of their experiments there was no evidence that age disproportionately affected VWM for conjunctions relative to individual features. In one experiment a significant age by condition interaction was found, although in this case it appeared to be due to specifically poor performance of the younger group in the shape only condition.

In summary, the growing literature on this topic suggests that there is no disproportionate age-effect on the ability to retain the combination of surface features in VWM relative to the features independently. This contrasts with the clear disproportionate effect of age on relational memory relative to item memory seen in LTM research (see Section 1.3). However, given the small number of experiments, the extent to which feature binding is generally age-invariant is unclear. The findings of Brown and Brockmole (2010) point to a potential boundary condition under which a reliable age-related binding deficit may be observed. Brown and Brockmole (2010) conducted two change detection experiments assessing the role of attention in VWM feature binding. Participants studied three objects, defined by colour and shape, and following a one second delay were presented with a single probe testing VWM for features or conjunctions. Experiment 1 compared a simple articulatory suppression condition to a more demanding backwards counting condition, whilst Experiment 2 compared simultaneous and sequential presentation of memory objects. Both manipulations led to greater disruption of binding performance relative to individual features; and this appeared to be true for both age-groups. However, a comparison of the two experiments yielded an interesting pattern of results. In Experiment 1 there was no evidence of an age-related binding deficit; that is, there was no significant interaction between age-group and memory condition. By contrast, in Experiment 2 there was evidence for an age-related binding deficit in the form of an age-group by memory condition interaction, with binding showing a larger age effect than individual features, in particular shape, alone. As Brown and Brockmole note, a key difference between the two experiments was the duration for which mem-

ory objects were presented. In Experiment 1 the memory array was presented for 900 ms, whereas for Experiment 2 this was increased to 1500 ms, due to sequential presentation in the more demanding experimental condition.

The surprising, and somewhat counter-intuitive, finding of Brown and Brockmole (2010) may reflect the temporal nature of feature binding in VWM. As we outline in greater detail below in Section 1.5, it may be that longer presentation durations support a more elaborative, resource demanding, form of binding that older adults may be less able to implement and benefit from. However, given that presentation time was not of interest to their experimental manipulations, in Chapter 3 we aim to directly assess the role of presentation time in the emergence of an age-related colour-shape binding deficit and the efficacy of feature binding more generally.

Clarifying further the effect of healthy ageing on the ability to bind the surface features of objects in VWM is important given that there appears to be a pronounced surface feature binding deficit associated with Alzheimer's disease (AD) (e.g. Della Sala, Parra, Fabi, Luzzi, & Abrahams, 2012; Parra, Abrahams, Fabi, et al., 2009; Parra, Abrahams, Logie, & Della Sala, 2010). In fact this feature binding deficit may emerge very early on during disease progression (Parra, Abrahams, Logie, Mendez, et al., 2010; Parra, Cubelli, & Della Sala, 2011). Parra, Abrahams, Logie, Mendez, et al. (2010) studied groups of participants with a rare form of familial AD caused by a mutation of the PSEN1 gene; one group who were largely asymptomatic at the time of testing and another who were displaying typical AD symptoms. These groups were compared to family members who do not carry the mutation. In a change detection task for colour, shape, and shape-colour binding, none of the groups differed greatly in their ability to detect changes to individual features. Unsurprisingly the symptomatic carriers showed a large deficit in the binding condition, in line with findings from sporadic AD (e.g., Parra, Abrahams, Fabi, et al., 2009). Crucially, the asymptomatic carriers also exhibited this deficit, albeit slightly attenuated. This is particularly striking given that, at the age they were assessed, these individuals would be expected to have approximately 10–15 years until meeting the diagnostic criteria for AD. Thus a feature binding deficit appears

to be a hallmark of early changes associated with AD and this has obvious implications for the assessment of at-risk older adults (Didic et al., 2011; Parra, 2014). Further, recent evidence suggests that a shape-colour binding deficit may be a specific hallmark of AD and not present in other dementias (Parkinson's, Lewy Body, Fronto-temporal, Vascular), suggesting the invaluable opportunity for differential diagnosis (Della Sala et al., 2012).

In summary, whilst there is overwhelming evidence for an associative deficit with healthy ageing, the effect of age on the ability to bind features in VWM is less clear. It has been suggested that age may impair the ability to maintain object-location associations in working memory, due to hippocampal dysfunction, however subsequent work on this has been less clear cut. Considering objects defined by combinations of colour and shape, older adults appear to do no worse at retaining the exact binding of features than would be predicted by their memory for the features individually. Nevertheless, the literature on feature binding in VWM is in its infancy and some studies have pointed towards situations where older adults may struggle to retain feature conjunctions. The findings of Cowan et al. suggest that mixing feature and binding changes in the same block of trials may specifically impair older adults' ability to notice binding changes. Further the findings of Brown and Brockmole suggest, rather counter-intuitively, that longer exposure durations of to-be-remembered items may lead to the appearance of an age-related binding deficit.

It is clear, therefore, that this question is in need of more data, in particular on the potential boundary conditions under which healthy older adults may exhibit a reliable feature binding deficit. The present work aims to contribute to our current understanding by testing some of these boundary conditions. As the focus is on feature binding, it is useful to outline, in some detail, theories of feature binding both in perception and working memory, as understanding of the latter has leaned heavily on work done in the former.

1.5 Theories of Feature Binding in Perception and Memory

The Binding Problem

In contrast to our experience of a perceptually coherent world there is considerable evidence that the brain is massively parallel, with separable processing streams for different feature dimensions. Early evidence from single cell recordings conducted on primates demonstrated that cortical streams responding to different feature dimensions, such as form, colour, and orientation, diverge very early on in the processing hierarchy (i.e. prior to visual cortex) and retain a great degree of separability in visual cortex and at higher levels (e.g. Hubel & Livingstone, 1987; Zeki, 1976). This animal work has been corroborated by neuropsychological studies of brain damaged patients with impaired colour perception but preserved discrimination of form (Heywood & Kentridge, 2003), and vice versa (Humphrey, Goodale, Jakobson, & Servos, 1994), as well as human neuroimaging studies of perception that show non-overlapping regions of activation associated with attention to different attributes (e.g. Cant & Goodale, 2007).

Of course anatomical division of labour does not necessarily mean that these different feature dimensions are psychologically separable; however there is plenty of psychophysical evidence for the binding problem (see, Wolfe & Cave, 1999, for a review). In his highly influential book, Garner (1974) outlines multiple sources of evidence for what he calls 'separable' and 'integral' feature dimensions. A crucial difference between these two classes is that, for separable dimensions it is possible to selectively attend to one dimension while ignoring irrelevant variation in another. For example, the speed and accuracy of sorting stimuli or making a classification on the basis of colour is not affected by irrelevant variation in shape (see also, Cant, Large, McCall, & Goodale, 2008). Whereas for integral features attention to one dimension *necessitates* attention to the other and thus the efficiency of sorting or classification suffers in the face of variation in the irrelevant dimension. Examples of such integral features include height and width, size and shape, as well as hue

and luminance (see also, Bae & Flombaum, 2013). A further source of evidence for the existence of separable dimensions comes from the study of *illusory conjunctions*, discussed in more detail below. When attention is diffuse it appears that shape and colour of different objects may be perceived as if they appeared together (Treisman & Schmidt, 1982). These anomalous perceptions would clearly not be predicted if the features were integral, as we would expect that identification of one object feature comes with identification of the others.

The physiological and psychophysical evidence contrasts with our perception of coherent integrated objects. Thus there is a *binding problem* that must be accounted for in order to explain how features that are apprehended in largely separate, parallel streams become unified to produce our coherent experience of the world (see, Treisman, 1996 for a review. Although see, Di Lollo, 2012 for a different view). There have been many attempts made to explain how this binding problem is resolved and I outline a select few below.

Binding in Perception

Feature integration theory

The Feature Integration Theory (FIT) of visual attention (Treisman & Gelade, 1980) is perhaps the best known attempt to explain how the binding problem is solved and has been hugely influential in the study of visual cognition over the past three-plus decades (see Quinlan, 2003, for a review). Treisman and colleagues (Treisman & Gelade, 1980; Treisman, 1977; Treisman, Sykes, & Gelade, 1977) initially proposed that features are identified pre-attentively, in parallel and form separate feature maps. These feature maps were said to be independent of a master map of locations, therefore spatial attention was required to shift between occupied locations in order to bind the features present into an object representation. In assessing this Treisman and Gelade (1980) reported a series of visual search experiments with both individual feature search and conjunction search conditions. In feature search conditions participants would signal whether or not they had detected a target feature (e.g. the colour blue or the letter S) embedded within a variable number of

distinct distractors (e.g. the letters T or X coloured in brown or green) as quickly as possible. In conjunction search conditions the task was to indicate whether a target pairing of features (e.g. a green T) was present within distractors made up of the same features but differently paired (e.g. brown Ts and green Xs). In line with the predictions of FIT, search functions relating the number of objects in the display to reaction time were flat when a feature target was present, suggesting fast parallel identification of individual features. When feature targets were not present there was an increasing, non-linear function suggesting some additional check that no targets were present. For conjunction search, however, the search functions were linear, as predicted by the serial allocation of attention, and the slope for target trials was around half that for non-target trials, suggesting that the search stopped when a target was found (i.e. it was self-terminating).

Another central proposition of the original FIT was that without focused attention features are essentially 'free floating', as there is no cross referencing across the different feature maps, and this can result in the perception of *illusory conjunctions*. Treisman and Schmidt (1982) conducted a series of experiments in which participants were briefly presented with two digits on either side of a set of coloured letters. Their participants' primary task was to report the identity of the digits and there were various secondary tasks that concerned the central coloured letters. Therefore, attention was diffuse and not enough time was given to serially scan across all items in the array. Crucially, Treisman and Schmidt found that participants regularly reported seeing conjunctions of colour and letter that were not conjoined in the array. This could not be attributed to memory failures as participants were also likely to claim that a pre-cued conjunction was present in a set when it was not. These illusory conjunctions did not appear depend on the distance between objects supporting the notion that features are not cross-referenced with the master location map without focussed attention. Further, these anomalous perceptions were often endorsed with high confidence, suggesting that they were actually experienced and not the product of guessing. This was taken by Treisman and Schmidt as strong evidence that without focussed attention features are free floating and can randomly

combine with features from other objects, regardless of spatial proximity, in order to reach conscious awareness.

Location uncertainty and binding

In contrast to the original findings of Treisman and Schmidt, several subsequent studies have found that, in fact, the spatial proximity of objects *does* affect the likelihood that their features will form illusory conjunctions (e.g., Cohen & Ivry, 1989, 1991; Prinzmetal & Keysar, 1989). This has led to the development of location uncertainty theories of feature binding in which there is no additional 'binding' stage in perception per se, but rather conjunction errors represent uncertainty in the location of the target object. For example, Ashby, Prinzmetal, Ivry, and Maddox (1996) outlined a mathematical model which proposed that, given a brief display, identity (e.g. letter) and colour are selected independently and the location of the features is inherently imprecise. Features are then combined on the basis of spatial proximity and, given the imprecise nature of location information, occasionally features from separate objects appear to come from the same object. Ashby and colleagues also created a formal version of FIT, which they called the 'random binding' model in which the proximity of items had no effect on the likelihood that their features would erroneously combine. In developing this model they noted that a version in which correct and incorrect binding were equally probable could already be ruled out on the basis of existing data; rather FIT is in line with the idea that on some proportion of trials when features are correctly identified binding takes place incorrectly. Ashby et al. assessed the ability of these models to explain performance in a partial report task where participants were briefly presented with a pair of coloured letters and had to report the identity and colour of a pre-specified target. The letter pairs could appear in one of 4 locations on the screen and the distance between the two letters was varied in order to distinguish the location uncertainty and random binding models. The location uncertainty models consistently provided better fit to individual level data relative to the random binding, FIT based, models as it was clear that the features of spatially closer objects were more likely to form illusory

conjunctions. Importantly both of these models outperformed a null model in which illusory conjunction reports were entirely due to guessing, suggesting that they do truly occur.

Further evidence for the role of spatial uncertainty and probabilistic sampling in perceptual feature binding was provided by Vul and Rich (2010). In their experiments participants were briefly presented with a circular array of coloured letters and were pre-cued to report both the identity and colour of a single target. The key manipulation was that the pre-cue appeared at various time intervals (0, 100, or 200 ms) before the onset of the study array thereby modulating the amount of time given to shift attention and consequently the internal uncertainty surrounding the location of the target. The colours and letters around the cued target were all different allowing the authors to assess the distribution of reported features around the target. Unsurprisingly, reporting errors were quite frequent when the amount of time given to shift attention was very short and reduced as more time was given. Crucially, however, the magnitude of error in letter and shape reports was completely uncorrelated; that is, errors in reporting colour appeared to be independent of errors in reporting identity. Vul and Rich replicated this finding in the temporal domain by presenting a rapid stream of letters with a cue co-occurring with the target item. Speeding up the rate of presentation increased the variability of reporting error but, again, these errors were independent. In line with the earlier proposals of Ashby et al., Vul and Rich propose that there is no special binding mechanism, per se, but rather when selection of a target (either in the spatial or temporal domain) is uncertain observers must sample from likely candidates (see also, Vul, Hanus, & Kanwisher, 2009). This sampling appears to take place independently between different feature dimensions (see also, Bundesen, Kyllingsbæk, & Larsen, 2003; Kyllingsbæk & Bundesen, 2007) leading to the lack of correlation between colour and shape reporting error. Therefore, according to this account correct binding occurs when attentional selection is precise enough to encompass a single object.

The difference between sampling accounts and FIT is subtle as according to both

solutions to the binding problem attention must be sufficiently precise in order to encompass a single object, otherwise binding errors may occur. However, the crucial difference is that in FIT an additional stage is required to serially shift spatial attention which serves to localise features. The findings of Ashby et al., Vul and Rich, and others (Bundesen et al., 2003; Johnston & Pashler, 1990; Kyllingsbæk & Bundesen, 2007) suggest that in fact features can be localised pre-attentively—albeit with some imprecision. Thus it seems that an additional stage for the localisation of features and formation of perceived feature conjunctions is unnecessary. Participants can localise features without focussed attention (with some uncertainty) and spatial proximity affects the likelihood that object features will combine. Thus a more parsimonious account is that of parallel selection of features within and between objects with correct feature binding being a limiting case in which attentional selection is precise enough (Vul & Rich, 2010) or if attention is biased towards certain objects within an array (Bundesen, 1990). The notion of independent sampling from multi-element arrays will prove useful later on in the thesis when discussing some of the recent findings regarding feature binding in VWM.

Binding in Visual Working Memory

The role of attention in maintaining feature bindings

Once the relevant information from a presented array has been selected and attended to does maintaining bound object representations in VWM require additional effort relative to maintaining individual features? In their classic paper, Luck and Vogel (1997) suggested that integrated object representations are stored automatically. By varying the number of items presented (set size) during a change detection task they found that participants performed almost perfectly when the number of items was small, up to approximately 3 or 4. However, performance greatly declined as set size increased beyond 4 items and this led them to suggest that VWM is limited in terms of the number of objects that can be stored (see also, Phillips, 1974; Sperling, 1960). Interestingly, Luck and Vogel (1997) also varied the number of features that each item was made up of. Participants had to maintain up to 4 features per item

(colour, orientation, size, continuity) in order to detect a change that could occur in any one feature dimension (other feature dimensions remained the same). Despite increasing the number of to-be-remembered features from 4 to 16, Luck and Vogel observed no clear change to performance (although see, Hardman & Cowan, 2015; Oberauer & Eichenberger, 2013). This they took to show that the unit of storage in VWM is the integrated, bound object and these are stored in VWM without cost.

Luck and Vogel noted that their results were also consistent with multiple independent memory stores for each feature dimension, rather than the integration of features into a single, unified representation. To address this they compared change detection performance when memory items consisted of a single colour to when participants had to remember two different colours per item. According to the parallel stores account adding more features from the same dimension to an object should result in poorer performance as the capacity for that feature dimension is exceeded. However, in contrast to this prediction Luck and Vogel found no cost associated with storing bi-coloured objects in VWM.

Wheeler and Treisman (2002) took issue with this finding, given that it ran contrary to the predictions of FIT which posits independent parallel feature maps. In their first two experiments they failed to replicate Luck and Vogel's findings as they clearly showed that change detection performance was greatly reduced when bi-coloured stimuli were presented relative to single coloured stimuli (see also, Delvenne & Bruyer, 2004; Olson & Jiang, 2002; Parra, Cubelli, & Della Sala, 2011). With the parallel feature stores account back in play, Wheeler and Treisman (2002) went on to test whether features were truly integrated in VWM. They noted that, as Luck and Vogel's task required the detection of *new* features at test, rather than a swap of feature combinations in the study array, the discrimination could have been based on feature memory only. For example, the observer may remember that the colour red was not in the memory array despite having no knowledge of the precise feature conjunctions presented. In order to assess whether observers had retained the correct conjunction in memory it is necessary that observers look for brand new *combinations* of previously studied features (cf. Treisman, 1977).

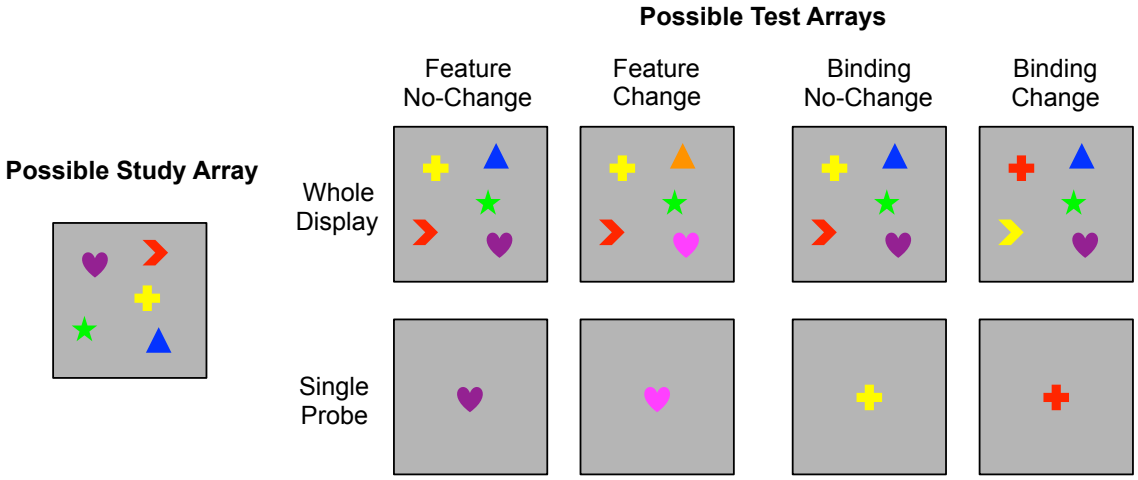

Figure 1.1: A possible study array and possible test arrays in the single probe and whole display change detection tasks used by Wheeler and Treisman (2002). Note that in the feature condition only colour changes are given, but there are additional conditions in which shape could change between study and test.

Across four separate experiments, Wheeler and Treisman assessed the binding of features in VWM. Two of these experiments assessed colour-location binding and two assessed colour-shape binding. For both the colour-location and colour-shape combinations single probe and whole display versions of the change detection task were compared (see Figure 1.1 for an example of colour-shape study and test arrays). It is important to note that the whole display task used here was not the conventional paradigm (introduced by Pashler, 1988) as on *change* trials two objects differed between study and test; in feature conditions this would mean the introduction of two brand new features, whereas in the binding conditions this would be a feature swap (e.g. two shapes swap colours. See Figure 1.1). Also in the colour-shape experiments location was rendered task irrelevant by shuffling items between study and test in the whole display task and by presenting the test item at the centre of the screen in the single probe task. The pattern of results was remarkably similar across the colour-location and colour-shape studies, so they will be discussed together.

In their whole display experiments, in which the same number of items were presented at study and test, Wheeler and Treisman found that binding change detection performance was much worse than individual feature change detection. However, in

their single probe experiments, where participants made a recognition judgement on a lone test item, change detection performance in the binding conditions was roughly equal to performance in the most difficult individual feature condition (colour for the colour-location experiment and shape for colour-shape). This close correspondence suggested to the authors that VWM stores bound objects, with capacity limited by the most difficult feature dimension. In order to explain the discrepancy between their two methods of probing VWM, Wheeler and Treisman proposed that, as the features appeared to be retained in parallel stores, there was an additional resource demanding function that served to link features from the same object and maintain that link in VWM. In the whole display paradigm it was suggested that this resource was diverted to processing the multiple test items, causing the bindings to disintegrate, whereas a single test object does not pose such an attentional demand allowing the links to remain.

The suggestion that maintaining feature bindings in VWM is a resource demanding process made sense in light of the supporting evidence for FIT (see above). However, it was quickly noted that Wheeler and Treisman (2002) did not directly test this claim. Given that FIT predicts a specific role of spatial attention in the formation of feature bindings several investigations have focused on whether distracting peripheral cues, that capture and shift spatial attention, can disrupt the maintenance of bound objects. However the majority of these have failed to demonstrate that the maintenance of shape-colour bindings is disproportionately disrupted by shifts of spatial attention (Delvenne, Cleeremans, & Laloyaux, 2010; Gajewski & Brockmole, 2006; J. S. Johnson, Hollingworth, & Luck, 2008; Yeh, Yang, & Chiu, 2005; although see Fougnie & Marois, 2009; Zokaei, Heider, & Husain, 2014).

There has also been much interest in the role of general attentional, or executive, resources in working memory binding given a modification to the multiple component model of WM (MCWM. Baddeley, 2007; Logie, 2011), namely the introduction of an 'episodic buffer'. Baddeley (2000) proposed the episodic buffer as a multi-modal store that serves to integrate information from modality specific buffers. The flow of information to this store was said to be under tight control from the

central executive component, consequently functions requiring additional binding of information were predicted to be reliant on a general attentional resource. Several studies have assessed this proposal with the introduction of demanding tasks, such as backwards counting in threes, to be performed concurrently with the change detection task. Studies assessing the binding of surface features have generally found no evidence that binding performance is disproportionately disrupted by concurrent tasks (R. J. Allen, Baddeley, & Hitch, 2006; R. J. Allen, Hitch, Mate, & Baddeley, 2012; C. C. Morey & Bieler, 2013), rather it appears that the maintenance of information in VWM in general is rather demanding of attention. Indeed these findings have led to a revision of the MCWM model in which the formation of bound representations takes place relatively automatically within the visuo-spatial sketchpad component before being passed on to the episodic buffer for conscious experience (Baddeley, Allen, & Hitch, 2011).

The question remains, then, why did Wheeler and Treisman (2002) find that a whole display probe disproportionately affected binding performance? Recent work has suggested that this may be an artifact of the single probe version of the task and that, in fact, this task may *overestimate* binding performance (Cowan, Blume, & Saults, 2013; Cowan, Saults, & Blume, 2014; H. Zhang, Xuan, Fu, & Pylyshyn, 2010). Given that the whole display version of the change detection task has fallen out of favour due to concerns regarding interference at test we reassess this methodological question in a series of experiments reported in Chapter 2. In addressing this open methodological issue, these experiments inform our approach in subsequent studies of feature binding in healthy older adults.

Although the extant literature does not appear to support a role for either sustained spatial attention or a more general form of executive control in the maintenance of feature combinations it has been suggested that this may be an exception rather than the rule. Mitchell, Johnson, Raye, Mather, and D'Esposito (2000) make the distinction between binding based on the products of early perceptual processing which are maintained in VWM without cost and 'memorial binding' which is presumed to operate when more time is given to encode stimuli into memory. Memorial

binding is said to rely on extra 'reflective' operations that serve to strengthen feature combinations by linking perceptual input to existing structures in long-term memory or by constantly attending to and 'refreshing' the information (see, M. K. Johnson, 1992). R. J. Allen et al. (2006) make a similar distinction between automatic binding which occurs when little time is given to study materials and active binding in which elaborative encoding strategies are used to strengthen feature associations. Whether or not there is a greater role for attentional resources when longer is given to study to-be-remembered items is at present unclear. However, there is some suggestion in the literature of a potential mediating role of presentation time. Elsley and Parmentier (2009) used a long exposure duration (2000 ms) relative to other studies (< 1000 ms) in their change detection experiments and found that concurrently maintaining words had a disproportionate effect on colour-shape binding performance. Given that Brown and Brockmole (2010) found an age-related binding deficit in an experiment in which participants were given longer to study memory items (1500 ms vs. 900 ms) it may be that older adults were less able to benefit from the extra time given to engage in a more active form of binding. We discuss this possibility further and address it directly in Chapter 3.

Is location special?

The initial proposal of FIT (Treisman & Gelade, 1980) was that during the pre-attentive phase of visual attention a master map of locations is established along with separate feature maps for colour, form, and so on. Spatial attention was said to then move between occupied locations to identify the features present and bind them together into object representations. However, subsequent work has shown that features appear to be intimately linked to spatial location early on in processing (i.e. pre-attentively) as accurately reporting feature identity from a briefly presented masked array appears to require accurate localisation (e.g. Johnston & Pashler, 1990; Nissen, 1985). Indeed as the work discussed above shows, it is possible to localise features, albeit with some uncertainty, without the serial allocation of attention (Bundesen et al., 2003; Kyllingsbæk & Bundesen, 2007; Vul & Rich,

2010). Regardless of the theoretical position it is clear that location as a feature occupies privileged status in visual attention and perception.

In the literature on VWM, however, the evidence concerning a possible privileged role for location is unclear. There is reason to believe that the links between object and location are rather fragile; indeed strict dependence on the initial presented spatial layout appears to be a crucial feature distinguishing a sensory (iconic) store from an abstracted VWM store (Phillips, 1974; Sperling, 1960). When a short delay (e.g. 100 ms) is interspersed between study and test screens, change detection performance is greatly disrupted by shuffling objects around relative to when the objects maintain their original locations. However, after a longer delay (beyond 1000 ms) this disruption to performance is greatly reduced (Logie, Brockmole, & Jaswal, 2011; Treisman & Zhang, 2006; Woodman, Vogel, & Luck, 2012). Also in recall tasks the probability that an object will be recalled in the incorrect position increases with greater retention time (Pertzov, Dong, Peich, & Husain, 2012). These findings suggest that representations in VWM become more abstracted and less dependent on their initial spatial location over time. This is also seen in the recent finding that, even though the use of location information would simplify memory search in the single probe change detection task, participants do not appear to use location to guide discrimination (Cowan et al., 2013; Gilchrist & Cowan, 2014).

That being said, there is evidence that when tasks explicitly require the maintenance of 'what was where' this form of binding is fairly robust, possibly more so than binding between surface features. Logie et al. (2011) report a series of experiments, using stimuli defined by colour, shape, and location, assessing the effect of varying a task-irrelevant feature on the maintenance of task-relevant bindings between the remaining features. When colour-shape conjunctions were relevant to the task irrelevant shuffling of locations between study and test was disruptive up to retention intervals longer than 1000 ms. However, when considering binding between a surface feature (e.g. colour) and location, task-irrelevant variation in the other feature dimension (e.g. shape) led to a smaller disruptive effect that had largely disappeared at intervals of around 500 ms. Thus when location is explicitly relevant

to a task it may retain some of its privileged status from early perceptual processing. Further evidence for the robustness of colour-location binding comes from Cowan et al. (2006) who used an auditory choice reaction time task to assess the role of focussed attention in this form of binding. As has been found with colour-shape pairings (see Section 1.5), performance was disrupted by the concurrent task, but the effect was no greater for the detection of colour-location binding changes relative to colour changes alone. Therefore, maintaining the combination of simple features and their locations with sufficient precision for recognition appears to be relatively cost free, although it may be that retaining precise object-location information for a recall task is more demanding of attention (Postma & De Haan, 1996).

The role of the medial temporal lobes, particularly the hippocampus, in binding together object and location in VWM has also received a great deal of attention given the well-established role of this structure in allocentric spatial processing (e.g., Ekstrom et al., 2003; O'Keefe & Nadel, 1978). Findings from studies of patients with MTL damage and neuroimaging studies of healthy participants have yielded mixed results. Olson, Page, Moore, Chatterjee, and Verfaellie (2006) presented a group of amnesic patients, with bilateral hippocampal damage (some with damage to other MTL regions), and healthy controls with a sequence of three objects on a 3×3 grid and following a short delay probed recognition memory for the objects, locations, or object-location pairings presented (cf. Mitchell, Johnson, Raye, Mather, & D'Esposito, 2000). The amnesic patients were proficient at recognising previously seen objects or locations but were specifically impaired at recognising conjunctions (see also, C. Finke et al., 2008; Pertzov et al., 2013). On the other hand, R. J. Allen, Vargha-Khadem, and Baddeley (2014) also presented stimuli on a 3×3 grid but instead used simple colours along with articulatory suppression. They found that their single case with selective hippocampal damage performed as well as (if not better than) control participants in recognising conjunctions. Further this single case was able to reconstruct colour-location pairs without difficulty (see also, Jeneson, Mauldin, & Squire, 2010).

The results from functional neuroimaging regarding the role of the MTL in

object-location binding have been as mixed as the findings of neuropsychological studies. Using fMRI Sala and Courtney (2007) found activation of ventral prefrontal cortex (PFC) associated with temporary maintenance of abstract colour patterns, whereas maintenance of spatial locations primarily activated dorsal PFC (as had been found previously; e.g., Haxby, Petit, Ungerleider, & Courtney, 2000). When participants maintained the conjunction of pattern and location, however, there did not appear to be any additional activity beyond recruitment of these areas of PFC. Crucially voxelwise analysis did not reveal any significant 'what was where' related activity in the medial temporal lobes (see also, Piekema, Rijpkema, Fernández, & Kessels, 2010). Nevertheless there have been other imaging studies that have suggested a role for the MTL in location binding. Piekema, Kessels, Mars, Petersson, and Fernández (2006) found that maintenance of sequentially presented letter-location conjunctions was associated with activity in the right hippocampus. Also, Mitchell, Johnson, Raye, and D'Esposito (2000) found activation of the left hippocampus when their younger participants were retaining object location correspondences.

Thus, there is some evidence for a role of the hippocampus in the formation and temporary retention of object-location conjunctions. Consequently we may expect an age-effect on object location binding given the pronounced senescent decline observed in this region (Raz & Rodrigue, 2006). However, as outlined above the evidence is far from conclusive and interpretation is made difficult by variation in methodology. For example, studies finding hippocampal involvement in object-location binding (Mitchell, Johnson, Raye, & D'Esposito, 2000; Olson et al., 2006; Piekema et al., 2006) or disruption by concurrent tasks (Elsley & Parmentier, 2009) tend to use highly nameable stimuli presented on a grid without articulatory suppression. It is possible that if participants are able to use verbal strategies the task becomes more like a measure of relational memory (e.g. car–top-right, umbrella–bottom-left), which is well-known to activate MTL structures (see, Shing et al., 2010, for a review), as opposed to a visual snapshot representation. The consideration of object-location binding is further complicated by evidence for different levels

of spatial representation; object-location bindings can be represented at a relatively categorical level of description (e.g. object A is above object B) or at a more fine grained coordinate level (Postma, Kessels, & van Asselen, 2008). The mode of presentation (simultaneous or sequential) and task requirements (recall or recognition) will likely modulate the contribution of these levels of representation. Further, there is a well known distinction between allocentric and egocentric spatial representation that may differentially contribute to measures of object-location memory (Baddeley, Jarrold, & Vargha-Khadem, 2011; Burgess, Maguire, & O'Keefe, 2002).

In Chapter 6 the evidence for age-related location binding deficit is critically reviewed and we report two experiments following on from the findings of Cowan et al. (2006). With simple conjunctions of colour and location we find evidence against a specific age-related deficit (see also, Read et al., 2016).

The above summary of our current understanding of feature binding in VWM and the effects of healthy ageing on this function point to a number of outstanding questions. However, before outlining how the present work aims to address some of these issues in more detail, it is important to outline a number of statistical and theoretical considerations when assessing the evidence for age-related binding deficits.

1.6 Statistical Considerations when Assessing Age-Group Interactions

In assessing the extant literature on age-related binding deficits it is important to take some statistical and methodological considerations into account. Implicit in the notion of an age-related binding deficit is that performance on tasks in which some extra binding is required should exhibit a greater effect of age than tasks that do not require binding (or pose less of a binding load). Therefore, the crucial statistical test is of an *age-group* $\times$ *condition interaction*. In some cases, however, tests of crucial interactions do not meet the conventional significance threshold and conclusions are based on separate analyses of group data (Mitchell, Johnson, Raye, Mather, &

D'Esposito, 2000; Fandakova, Sander, Werkle-Bergner, & Shing, 2014). In other cases the relevant information needed to assess the evidence for these interactions is not presented (Borg, Leroy, Favre, Laurent, & Thomas-Antérion, 2011; Mitchell, Johnson, Raye, & D'Esposito, 2000). For any demonstration of an age-related binding deficit to be convincing it must be supported by the relevant interaction test. Chapter 5 discusses these issues in more detail in relation to the common suggestion that older adults have a specific VWM deficit for retaining what was where (i.e. object-location conjunctions).

However, there are less obvious aspects surrounding the evaluation of age × condition interactions that have the potential to introduce far greater bias into this literature and related fields.

Assessing Age-Group Interactions in Recognition Tasks

In his discussion of methodological issues in cognitive ageing research, Salthouse (2000) notes that interpretation of age by condition interactions poses a particular problem. It is well known in statistics that non-cross over interactions can be transformed away with a non-linear monotonic transformation. Consequently, non-interactions at the level of the psychological construct (e.g. familiarity signal elicited by a previously seen item) may appear as interactions at the measurement level (e.g. proportion correct, reaction time) or vice versa (Salthouse, 2000). This problem is especially apparent when considering age-differences in same-or-different recognition tasks. Change detection data are inherently binary (correct or incorrect; 1 or 0) and the crucial measure of interest is often the probability of a correct response.

It is common practice, in ageing studies and the recognition literature more generally, to calculate the proportion of correct responses for each participant in each condition and conduct an analysis of variance (ANOVA) on the resulting scores. This presents several problems; firstly proportions are bounded between 0 and 1 and the number of distinct values the estimates can take is determined by the number of trials in the experimental design. If the number of trials is different across conditions—and the probability of a correct response is estimated with different levels of precision—

this crucial information is lost via aggregation. Further, binomial data violate the homogeneity of variance assumption[2] as the sample variance is highest at proportions around 0.5 and decreases as proportions approach 0 or 1. The consequence of this is that differences in proportions in the range of 0.5 to 0.7 matter less than equivalent differences in the range of 0.7 to 0.9 (Jaeger, 2008). Given that older adults tend to perform at a lower level overall relative to younger adults this presents a particular problem in experimental cognitive ageing research.

These problems are well known, particularly in psycholinguistics, and recently there have been increasing calls to move towards the use of generalised linear models (GLMs) in the analysis of categorical data (e.g., Agresti, 2002; Bolker et al., 2009; Jaeger, 2008). GLMs allow non-normal response variables to be modelled as a linear combination of predictors via a link function (Nelder & Wedderburn, 1972). A principled approach to analysing binomial data is to use logistic regression (logit link function) in which the log odds of a correct response is modelled as a linear function of the factors in the experimental design (Dixon, 2008)[3]. As shown in Figure 1.2A this transformation captures the fact that differences in proportions near 0.5 matter less than differences near either end of the scale. This logit transform can conceptually be thought of in terms of relative 'response strength' for the correct and incorrect answers, with the ratio of response strengths affected by the design factors (Dixon, 2008).

Figure 1.2 highlights the specific concern when assessing age-differences with categorical data. As noted above, proportions towards the middle of the 0–1 scale are more variable than proportions at the boundaries and, as recognition performance is typically lower, older groups tend to occupy the range of performance where differences matter less. Panel A provides an example in which on the logs odds (or response strength) scale (x axis) there are two orthogonal main effects, one of age and

[2]For a Binomial random variable the sample variance depends on the probability of success, p: $\mathrm{Var}(p) = \frac{p(1-p)}{n}$.

[3]The rationale behind the use of logistic regression can be explained as follows: Rather than modelling the probability of a correct response, p, focus is shifted to the odds of a correct response, $p/(1-p)$. When discussing odds it is common to talk in terms of multiplicative effects (e.g. x is twice as likely as y), so taking the natural logarithm allows us to talk of additive effects across all real numbers (Dixon, 2008). The link function used is the logit, $\mathrm{logit}(x) = \log(x/(1-x))$, and the inverse link function is the logistic, $\mathrm{logistic}(x) = 1/(1+e^{-x})$ (see Figure 1.2A).

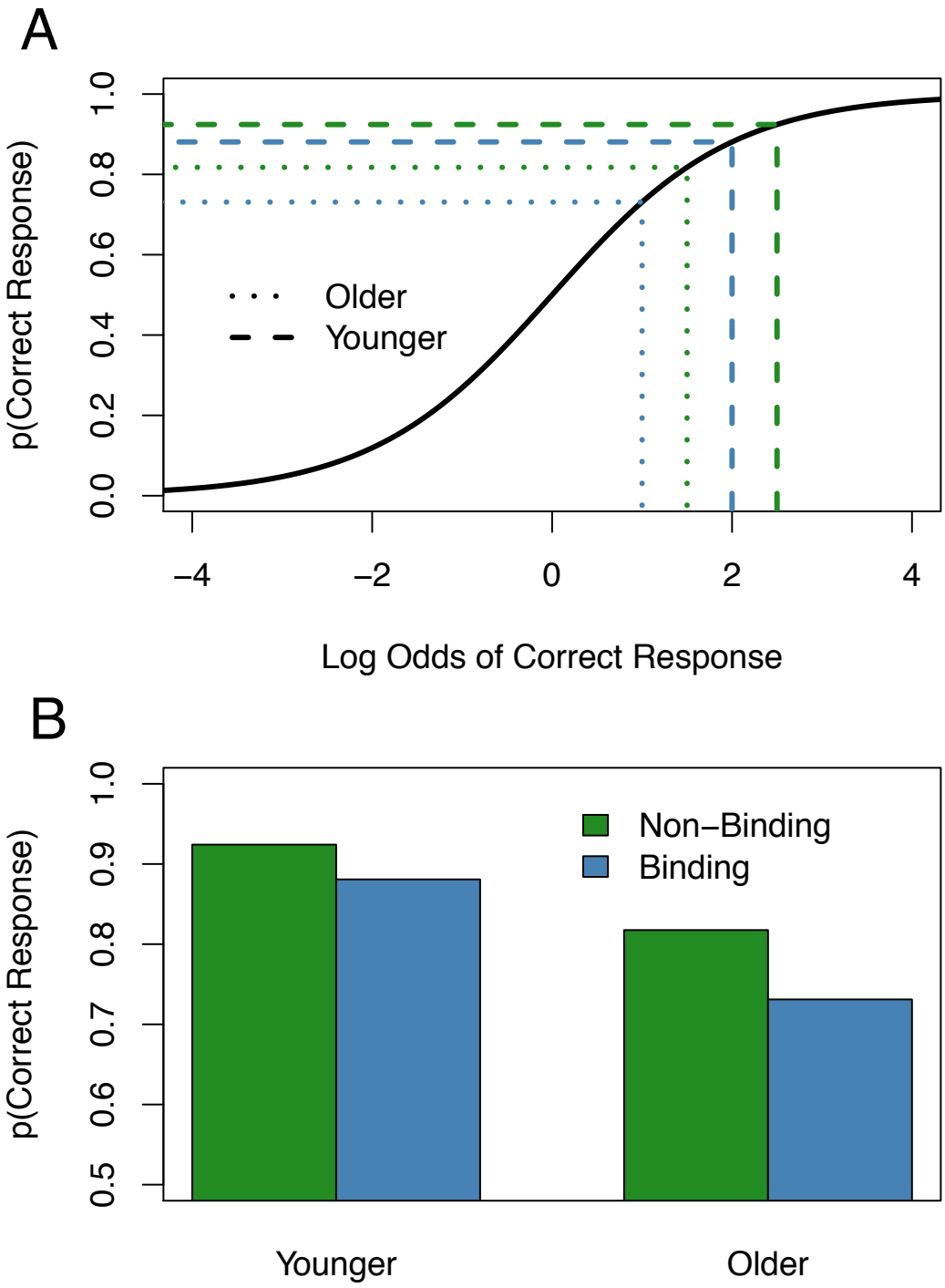

Figure 1.2: Illustration of a potential problem in assessing age effects with binary data. Panel A shows the logistic function that relates the log odds of a correct response to the probability of a correct response. This is the inverse link function used in logit modelling. The lines demonstrate that a non-interaction on one scale can appear as an interaction on another. Panel B shows this spurious interaction more clearly.

one of condition (in this example binding and non-binding tasks), and no interaction. Through the 'S' shaped transformation a spurious interaction is produced on the scale of measurement (probability of a correct response, y axis). Figure 1.2B shows this more clearly. With a large enough sample a plot like this would certainly result in a significant age by condition interaction and the (incorrect) conclusion that binding is disproportionately affected by age. Dixon (2008) present simulations of binomial data with two orthogonal main effects on the log odds scale—which, for our purpose, can be thought of as age-group and condition—and showed that the use of a normal model (e.g. ANOVA, linear regression) on aggregated proportions was susceptible to giving spurious evidence for the two-way interaction (i.e. a large type one error rate). The evidence for the interaction increased as the magnitude of the main effects was increased, this makes sense given Figure 1.2 as increasing the effect of age, for example, would place older adults further down the curve making the difference between the two conditions appear larger. Dixon (2008) concludes that "[...] as a default assumption in the absence of more theoretically guided choices, the logistic model is superior to the normal model." (pp. 451).

Consequently our approach to analysing raw (correct/ incorrect) data is to use a hierarchical logistic regression model to avoid the possibility of spurious interactions arising from the use of an incorrect data model. The details of this analysis are outlined in Chapter 2. This 'model free' analysis will be accompanied by analysis of measures that aim to separate the contribution of sensitivity and bias to task performance. Separating out sensitivity (or discriminability) from response bias requires a model of how the recognition judgement is achieved (Snodgrass & Corwin, 1988). Crucially different models make different assumptions about how differences in sensitivity should affect the rate of hits and false-alarms and choosing between measures can greatly effect conclusions.

Measures of Sensitivity/ Discriminability

In attempting to isolate the contribution of sensitivity and bias to recognition (or detection) performance, two broad classes of model prevail;

1. Those based on *detection theory* assume that items are judged on a single decision variable (e.g. familiarity) with old items tending to elicit higher values on this variable relative to new ones (Macmillan & Creelman, 2005). Measures derived from detection theory differ in the underlying distributions that old and new items are selected from, with the most common variant (d') assuming that the underlying distributions are Gaussian with equal variance;

2. Those based on *threshold* models of recognition assume discrete states where the observer either has the relevant information to make the discrimination or is in a state of ignorance and must guess between two alternatives (Snodgrass & Corwin, 1988). The most popular measure derived from this conception assumes that the probability the observer has the relevant information in memory is the same for old and new trials (two-high threshold; Pr).

There is an additional measure of sensitivity that is commonly used in the literature that does not have a clear underlying model. A' was derived to estimate the area under the receiver operating characteristic (ROC) curve (or isosensitivity curve) from a single hit and false alarm pair (Pollack & Norman, 1964). While it was derived without explicit reference to underlying distributions, and is thus commonly referred to as 'non-parametric', it has since become clear that A' *does* make distributional assumptions (e.g., Macmillan & Creelman, 1996; Pastore, Crawley, Berens, & Skelly, 2003). These measures are outlined in greater detail in the Introduction to Chapter 8.

While it is often treated as such, the choice between these measures is not arbitrary. Ultimately the shape of the ROC curve determines which model is more appropriate for a given task; a linear ROC curve with a slope of 1 is more consistent with a two-high threshold model whereas a symmetrical non-linear curve is more consistent with the Gaussian equal variance model underlying d' (Swets, 1986b). Of course it is not possible to gauge the shape of the ROC curve with a single hit and false alarm pair, therefore the choice between measures must be informed, wherever possible, by the literature (Swets, 1986a). It is worth noting that for the change detection task with highly discriminable stimuli the empirical ROC curve is consis-

tent with a threshold model of recognition (see, Rouder et al., 2008). Inappropriate choice of recognition measure has been shown, via simulation, to increase the likelihood of type I error when comparing two conditions (Rotello, Masson, & Verde, 2008; Schooler & Shiffrin, 2005). However, the effect of choosing between recognition measures on type one error rate for tests of *interactions* has, to our knowledge, not been assessed.

There is some suggestion in the literature on feature binding in VWM that the choice between measures *does* affect conclusions. Effect size estimates (e.g. partial eta squared) for age-group by condition interactions are generally larger when using A' relative to, say, proportion correct (e.g. Brown & Brockmole, 2010). As well as changes in magnitude, changes in significance have also been reported. Isella et al. (2015) reported a replication study of Brockmole et al. (2008) and found a significant age-group by condition interaction in an analysis of A', consistent with a larger effect of age for feature bindings relative to individual features. In their supplementary material Isella et al. (2015) report analysis of proportion correct with no hint of the crucial age-group by condition interaction.

While the authors explained the A' effect away—arguing that the contrast between shape and binding did not survive correction for multiple comparisons (pp. 40–41)—this clearly shows that the choice of measure can greatly affect the (potential) conclusions of studies of age-related binding deficits (R. J. Allen et al., 2012, also found that different measures led to different patterns of interaction effects). The extent to which this issue has introduced a bias in the literature on age-related feature binding deficits is unclear. Thus in Chapter 8 a series of simulation studies are reported that assess the effect of choice of measure on the type I error rate for tests of group by condition interactions.

In addition to analysis of commonly used measures of sensitivity we also make use of simple processing models derived from the slots conception of working memory (Cowan, 2001; Cowan & Rouder, 2009) to further probe the effect of age on VWM. The measures taken from this view of VWM are conceptually similar to those from the threshold theory of recognition and are outlined in greater detail in the next

chapter. In Chapter 7 these models are used to further explore the data collected in Chapters 4 and 5 to assess the contribution of capacity, guessing strategy, and the frequency of lapses of attention to age-differences in change detection performance (R. D. Morey, 2011; Rouder et al., 2008; Rouder, Morey, Morey, & Cowan, 2011).

1.7 Overview of the Present Work

There is overwhelming evidence that healthy ageing is accompanied by an associative deficit for disparate items (Old & Naveh-Benjamin, 2008a). This has led to interest in the possibility that binding deficits underlie the decline of VWM seen across the lifespan (Brockmole et al., 2008; W. Johnson et al., 2010). The literature, thus far, has largely suggested that the ability to maintain conjunctions of colour and shape is no more affected by age than the ability to maintain individual features. This has potential practical implications given the pronounced deficit observed in sporadic and familial Alzheimer's disease (Parra, Abrahams, Fabi, et al., 2009; Parra, Abrahams, Logie, Mendez, et al., 2010). However, previous work also suggests potential boundary conditions under which reliable age-related feature binding deficits may occur. It is these conditions that form the focus of the present work.

The first potential boundary condition we assess is that of lengthy exposure durations. Brown and Brockmole (2010) found evidence of a specific age-related colour-shape binding deficit when participants were given longer to study memory objects and, as discussed above, it is possible that this reflects the use of a more elaborative form of binding by the younger adults that older adults were less able to benefit from. In assessing this question directly Chapter 3 reports evidence against the suggestion that presentation time differentially affects the performance of younger and older adults.

As previously outlined, Cowan et al. (2006) found an age-related binding deficit when changes to colour-location pairing were mixed with changes to colour alone, but not when these different types of trial were presented in separate blocks. This finding may reveal a role for test salience in older adults' ability to detect binding changes, however the limited literature on this is unclear (T. Chen & Naveh-Benjamin, 2012).

In Chapters 4 and 5 a series of experiments are reported examining the effect of mixing versus blocking trials on older adults' change detection performance. Across four conditions with almost 200 participants we fail to show any effect of mixing trial types on performance. This is possibly the strongest demonstration so far that healthy ageing does not differentially affect the ability to form temporary feature conjunctions in VWM.

Further in this series of experiments comparing mixed and blocked trials we also included experiments assessing colour-location binding (Chapter 5). Thus we are able to compare the binding of surface features to location binding using more or less identical paradigms, which has not been done previously, to ask whether there is evidence that binding to location is a specific problem for healthy older adults. The pattern of results in our location experiments largely matches that found in our experiments on colour-shape binding, suggesting that, at least for simple features, location is not a particular problem for healthy older adults. Research into object-location binding is particularly complicated and slight variations in methodology are likely to vastly change patterns of results (see Section 1.5), therefore in discussing our findings we make a number of concrete suggestions for further research.

Our repeated inability to find a disproportionate age-effect on binding in VWM stands in stark contrast to the literature reporting age-related decline in memory for associations between items. The stimuli used in the studies of associative memory tend to be complex and ecologically valid, such as pictures of faces and scenes, whereas the stimuli used in research on VWM are comparatively simple. It has been suggested that this may underlie the discrepant findings between the two literatures (T. Chen & Naveh-Benjamin, 2012). However, binding features within objects and binding the relation between distinct objects are said to be two 'levels of binding' (Zimmer, Mecklinger, & Lindenberger, 2006; Zimmer & Ecker, 2010) that may be differentially affected by healthy ageing. Thus in Chapter 6 we attempt to directly contrast different forms of memory binding using simple stimuli defined by colour and shape.

However, prior to reporting the results of the ageing studies it is important to

address a methodological issue present in the field. As outlined above, Wheeler and Treisman (2002) showed that when participants were tested with a whole display probe binding performance appeared to be much lower than when a single probe was used. Subsequent work has suggested that the single probe task overestimates binding performance, thus the comparison of single probe and whole display procedures clearly needs reassessing. We use processing models derived from the slots conception of VWM (outlined in the next chapter) to better compare the two tasks. Across three experiments we find evidence that a whole display test results in little-to-no interference at test, relative to a single probe. These findings inform our methodological approach in our studies of healthy ageing. Further, in Chapter 7 the processing models developed in Chapter 2 are extended in an attempt to explore the contribution of capacity and lapses of attention to age-related working memory decline.

Finally, in Chapter 8 we present a number of simulation studies assessing the effect of choosing between different measures of sensitivity (or discriminability) on the likelihood of type I error for tests of age-group by condition interactions. To our knowledge this vital question has not yet been addressed and in doing so we find that choice of an incorrect task model can lead to out of control error rates. So much so that it is perhaps remarkable that so few studies of ageing and feature binding in VWM have found the crucial age by condition interaction. With this work and our assessment of the potential boundary questions in mind, we are better placed to evaluate the strength of evidence for a VWM feature binding deficit in healthy old age.

Chapter 2

Probing Visual Working Memory

2.1 Introduction

Prior to outlining our studies of healthy ageing, it is important to attempt to address a methodological issue in the field. Namely, how should short-term recognition memory be tested in the change detection task to best assess the retention of features and objects? Should the probe contain the same number of items as the original memory set or should a single item be judged? The findings of Wheeler and Treisman (2002) suggest that a single probe is best for assessing storage of feature bindings as a whole display probe appeared to cause binding specific interference at test in younger adults (see also, Kondo & Saiki, 2012; Yeh et al., 2005; although see, J. S. Johnson et al., 2008). Nevertheless, as outlined in Chapter 1, recent evidence suggests that the single probe task *overestimates* binding performance (Cowan et al., 2013; H. Zhang et al., 2010). Therefore, the apparent binding-specific whole display interference effect may be an artifact, caused by comparing the whole display task to a task which unfairly favours binding. In the present chapter we reassess the difference between single probe and whole display approaches to probing VWM. This section leans heavily on simple processing models derived from the slots account of VWM and the general philosophy underlying these models is outlined in detail below.

Estimating the number of items in VWM with change detection

As outlined in the Introduction chapter, the change detection task has proven extremely useful in studying the limits of VWM. Observers are presented with a variable number of items—varying on one or many feature dimensions—and following a short delay (usually 1 second) are given a *same*-or-*different* recognition test. Many versions of this task have been developed, primarily differing in the way that VWM is probed. Figure 2.1 shows the standard single probe version of the change detection task (left panel) in which a single item is tested at a previously occupied location along with the whole display version in which the array is re-presented with the possibility that one item has changed (right panel). The simplicity of the change detection task has led to the development of processing models that allow researchers to estimate, across a series of trials, the number of items an observer could retain in VWM and use to perform the discrimination. These processing models propose an item limit to VWM, with an observer able to retain k items, and when the number of items presented, N (set size), exceeds the capacity of the observer only k items are stored in VWM *with no information held about the remaining objects*. In addition, these models make a high-threshold assumption, which is that if the observer has the relevant information in VWM it will always be sufficient for the observer to detect a change (or no-change)[1]. As we discuss later, this assumption is highly controversial but appears reasonable when categorically distinct features are used (Donkin, Nosofsky, Gold, & Shiffrin, 2013; Donkin, Tran, & Nosofsky, 2014; Rouder et al., 2008). It is also crucial that these stimulus features are sampled without replacement as the opportunity to group items together may artificially inflate estimates of k from change detection tasks (Cowan, 2001).

As Rouder et al. (2011) outline, different processing models for estimating k are appropriate for different formulations of the change detection task. Consider the single probe change detection task, as depicted in Figure 2.1. According to

[1]Here we use *change* and *no-change* to refer to types of trials in the change detection task and *different* and *same* to refer to responses given by participants.

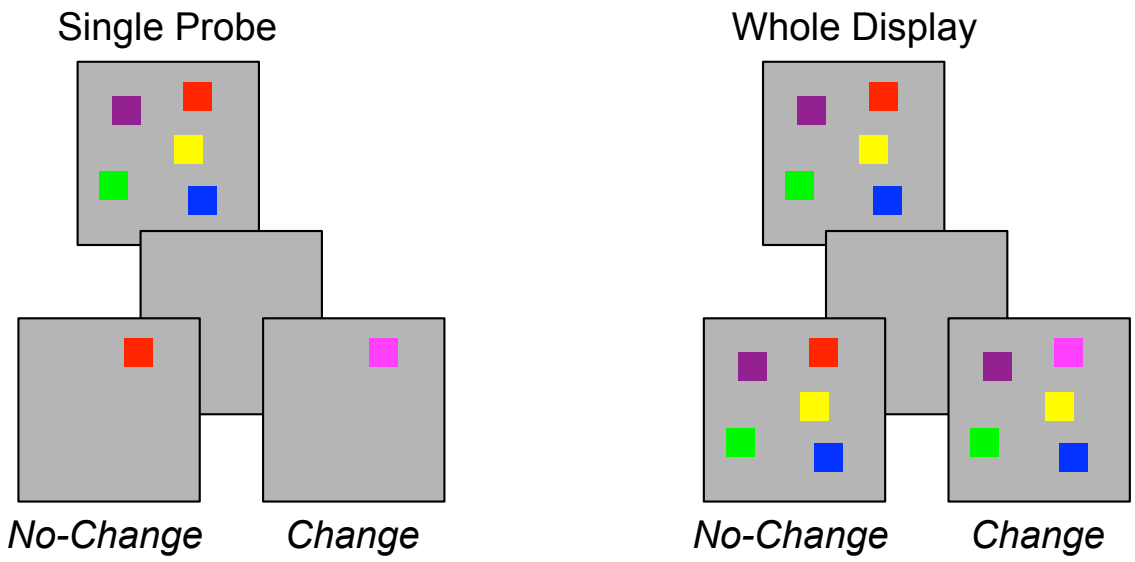

Figure 2.1: The standard single probe and whole display change detection tasks. The appropriate models for estimating the number of items in VWM for each of these tasks are described in the text.

the slots notion of VWM on a change trial the probability that the item at the probed location is in VWM, and hence that the change will be noticed, is given by $d = \min(k/N, 1)$. However, when the probe item is outside of VWM (on $1 - d$ proportion of trials) it is possible that the observer will guess correctly that a change has occurred with some probability, g. Therefore, the expected hit rate for an observer in this task, that is the probability that they will correctly identify a change, is given by $h = d + (1 - d)g$. Similar logic results in a prediction for false-alarms on no-change trials in this task, $f = (1 - d)g$. Cowan (2001) proposed this processing model and solved for, $\hat{k} = N(\hat{h} - \hat{f})$. This will provide an estimate of capacity provided: $k \leq N$ and $\hat{h} \geq \hat{f}$ (see, Rouder et al., 2011).

Similarly to the single probe task, in the whole display task, shown in Figure 2.1, a change will be detected if the changed item is present in VWM or the observer may correctly guess that a change occurred. Therefore, for the whole display task the prediction for hit rate is identical to that for the single probe task, $h = d + (1 - d)g$. When no-change has occurred in the whole display task a false-alarm can only arise as the requisite information was outside of VWM and the observer guessed *different*. Consequently, $f = g$. Pashler outlined this model in 1988 which results in the following equation for capacity, $\hat{k} = N\left(\frac{\hat{h}-\hat{f}}{1-\hat{f}}\right)$. This will provide an estimate

of capacity provided: $k \leq N$, $\hat{h} \geq \hat{f}$, and $\hat{f} < 1$ (see, Rouder et al., 2011).

Although it is rarely mentioned when discussing these models (although see, R. D. Morey, 2011) they are part of a class of models that have been hugely influential in cognitive psychology, multinomial processing tree models (Erdfelder et al., 2009; Riefer & Batchelder, 1988). The availability of the simple equations to estimate the number of items in VWM outlined above has in no small part led to the explosion of research on VWM capacity and its correlates (e.g., Cowan et al., 2005; M. K. Johnson et al., 2013; Van Snellenberg, Conway, Spicer, Read, & Smith, 2014; Vogel & Machizawa, 2004). Crucially, different models are logically appropriate for different versions of the change detection task and the potential objections to the findings of Wheeler and Treisman (2002) become clear when the appropriate processing accounts of their tasks are outlined, as we do below.

First, it is important to distinguish these measurement models from more specified accounts of WM processes (such as the serial order in a box model of complex span, Oberauer, Lewandowsky, Farrell, Jarrold, & Greaves, 2012). Given the assumptions of the simple processing models outlined here, which appear reasonable for supra-threshold stimuli, researchers can measure ostensibly important cognitive parameters (e.g. the number of items that could be retained). However, the process(es) underlying the obtained parameter values are left unspecified. They could reflect a single cognitive bottleneck resulting in a finite number of stimuli that can be retained or they could reflect the output of several, independent cognitive processes that contribute to the estimated capacity limit (Cowan et al., 2014; Logie, 2011). Here we take a pragmatic approach to using these processing models allowing us to better compare methods of probing VWM and this will hopefully inform future attempts to better specify the mechanisms underlying the model parameters.

Processing models for Wheeler and Treisman's tasks

The formulae of Cowan (2001) and Pashler (1988) were developed for specific versions of the change detection task (see Figure 2.1). Wheeler and Treisman (2002) addressed questions that could not be adequately assessed using these standard

tasks. In order to test that observers had correctly bound object features in VWM they used 'swap detection' where a single probe could be made up of two features taken from different memory items (e.g. a red square and blue circle become a red circle probe) or a whole display could contain two items that had swapped a single feature dimension (e.g. two shapes swap colours). In order to match the two changes in the whole display binding condition, conditions assessing feature memory also had to include two changes (i.e. features not in the studied set). Finally, to assess whether colour and shape were bound together, and not merely via location, location was rendered irrelevant in their studies by using a central location for the single probe and shuffling items in the whole display. Thus there were quite a few departures from the standard change detection tasks. Following the logic underlying the slots conception of VWM we find that *three* different processing models are appropriate for the change detection tasks used by Wheeler and Treisman (2002).

Single probe - Individual features

Cowan et al. (2013) point out that the measure proposed by Cowan (2001) is only appropriate when the single probe item is presented in its previously occupied location. In this case the observer can use their knowledge of location to restrict memory search to the single item which, according to the high-threshold notion of slots, is either in memory or not. However, Wheeler and Treisman (Experiment 4B), and subsequent studies of shape-colour binding in VWM (e.g. R. J. Allen et al., 2006), presented their probe item at the centre of the screen in order to render location uninformative. Cowan et al. (2013) provide the appropriate processing model for this version of the task where VWM for individual features is probed. In this case if the probe has been selected from the initial memory set (i.e. no-change has occurred), the observer detects this if this probe item is in VWM which occurs with the rate, $d = \min(k/N, 1)$, where N denotes the number of objects presented (set size) and k denotes the number of items (in this case, individual features) the observer can retain. If the probe item is outside VWM the observer must guess whether a change occurred or not. The probability of incorrectly guessing different

is given by g. Therefore the probability of incorrectly responding *different* for this task is given by,

$$f = (1 - d)g.$$

When a change has occurred an incorrect *same* response (i.e. a miss) can only arise due to guessing. Therefore,

$$1 - h = 1 - g,$$

provided that $k < N$ otherwise the observer would not be expected to miss any changes. Cowan et al. (2013) combined the above equations and solved for,

$$\hat{k} = N\left(\frac{\hat{h} - \hat{f}}{\hat{h}}\right), \tag{2.1}$$

which provides an estimate of k provided that $k \leq N$, $\hat{h} \geq \hat{f}$, and $\hat{h} > 0$.

Single probe - Binding

H. Zhang et al. (2010) note that in the single probe binding condition, as a change involves two objects donating features to the recombined probe, the observer will detect the change if they have either or both of the changed items in VWM (see also, Cowan et al., 2013). This occurs with the probability[2],

$$c = 1 - \frac{(N - k)(N - k - 1)}{N(N - 1)},$$

provided that $k < N - 1$, otherwise the adequate information would certainly be in VWM. Of course, here k denotes the number of bound objects the observer can retain. If the requisite information is outside of VWM it is possible that observers correctly guess that a binding change occurred. With this in mind the probability of correctly identifying a change in the single probe binding task is,

$$h = c + (1 - c)g.$$

When no binding change occurs the observer notices this when the probed item is in VWM, which occurs at a rate of $d = \min(k/N, 1)$. This leaves the probability

[2]To maintain consistency with other models reported here we have made a number of changes to the way in which this model is presented relative to the original paper. However, the fundamental features of the model proposed by H. Zhang et al. (2010) remain unchanged.

of incorrectly responding *different*,

$$f = (1 - d)g.$$

H. Zhang et al. (2010) combine and solve for,

$$\hat{k} = \frac{N(1-\hat{f}) + N - 1 - \sqrt{(N(1-\hat{f}) + N - 1)^2 - 4N(N-1)(1-\hat{f}+\hat{h}-1)}}{2}, \tag{2.2}$$

which provides an estimate of k in this condition if $k \leq N - 1$ and $\hat{h} \geq \hat{f}$.

Whole display - Individual features and binding

As we outlined above, change trials in the whole display task used by Wheeler and Treisman involved *two* changes in both individual feature and binding conditions. Therefore, the same general processing model is appropriate for both conditions. In this case correct identification of a change occurs when the observer has *either or both* of the changed items in VWM. The probability of this occurring is given by c (see above), with k referring to the number of features or bound objects held, depending on the condition. Again, if the relevant information is not in VWM the observer may guess correctly. Therefore, the probability that participants will correctly identify a change in this whole display task is,

$$h = c + (1 - c)g.$$

An incorrect *different* response when no-change has occurred can only arise due to guessing as the array size exceeded k, otherwise no false-alarms are made. Thus,

$$f = g,$$

as long as $k < N - 1$. With the above equations we combine and solve for,

$$\hat{k} = \frac{2N(\hat{f}-1) + 1 - \hat{f} + \sqrt{1-\hat{f}}\sqrt{4N(\hat{h}-1-\hat{h}N+N) + 1 - \hat{f}}}{2(\hat{f}-1)}, \tag{2.3}$$

which gives an estimate of k if $k \leq N$, $\hat{h} \geq \hat{f}$, and $\hat{f} < 1$.

As alluded to above the discrete state assumption is controversial and debate is very much on going (Luck & Vogel, 2013; Ma, Husain, & Bays, 2014; Suchow,

Fougnie, Brady, & Alvarez, 2014). However, when distinct stimuli are used the threshold assumption appears to give a reasonable account of performance in the change detection task (Donkin et al., 2013, 2014; Rouder et al., 2008) and results in useful measurement models for the present circumstance.

2.2 Experiment 1 – Reassessing Whole Display Interference

As we have outlined, the change detection tasks used by Wheeler and Treisman (2002) pose completely different demands in terms of the number of items needed in VWM to obtain equivalent levels of performance. Specifically the single probe task may overestimate performance for bindings relative to individual features when proportion correct is used as the outcome measure. In the present experiment we aimed to recreate Wheeler and Treisman (2002)'s key findings using proportion correct and apply the simple processing models described above to better compare the two methods of probing VWM. Further, unlike Wheeler and Treisman (2002) we compare the two testing methods within the same group of participants, giving us greater power to detect effects where they occur.

Methods

Participants

Twenty-four participants (aged 18–30, 14 females) were recruited from the student community of the University of Edinburgh. Each participant received payment of £ 5 for the 45 minute testing session.

Stimuli

Items were presented on a grey background on a 20" LCD computer screen. Memory arrays consisted of either 4 or 6 coloured shapes[3] each subtending 1.3° of visual angle

[3]Wheeler and Treisman (2002) used an additional set size, 2, however we decided to omit this condition firstly to avoid ceiling performance and secondly to avoid making the experiment too

at an approximate viewing distance of 60 cm (unconstrained). Each object appeared in one of 8 randomly selected locations on a 3×3 grid, measuring $6.7° \times 6.7°$, surrounding the centre of the screen, (the middle location was not used in the memory array). Objects in the memory array were constructed by selecting 4 or 6 colours and shapes randomly without replacement from master sets of 8 features. The 8 possible colours were brown, pink, orange, purple, green, blue, red, and yellow, and the 8 possible shapes were arch, hourglass, plus, star, circle, flag, diamond, and chevron. In the single probe condition the test array consisted of a single item presented at the centre of the screen, rendering location uninformative. In the whole display condition the test array contained the same number of items as the initial memory array. To render location uninformative in this condition items were shuffled within the original locations randomly between study and test (cf. Wheeler & Treisman, 2002).

Design and Procedure

Participants initiated each trial by pressing the space-bar on the keyboard. The trial began with a fixation cross presented for 500 ms followed by a 250 ms blank screen before the memory array was presented. The memory array remained visible for 500 ms and was replaced by a blank screen for a 1000 ms retention interval. The test array then appeared and remained visible until the participant made a response.

The experiment was split into 6 blocks combining the two versions of the change detection task (single probe and whole display) and three memory conditions (colour only, shape only, and colour-shape binding), as shown in Figure 2.2. In the single probe task participants were instructed that half of the time the test object would have been in the memory array and the other half of the time it would have changed. In individual feature conditions (colour or shape only) they were instructed that a change would involve the introduction of a brand new feature (colour or shape) that had not appeared in the original set, and the task-irrelevant feature would not change. In the binding condition they were told that a change trial would involve

long given the manipulation of task type within participants.

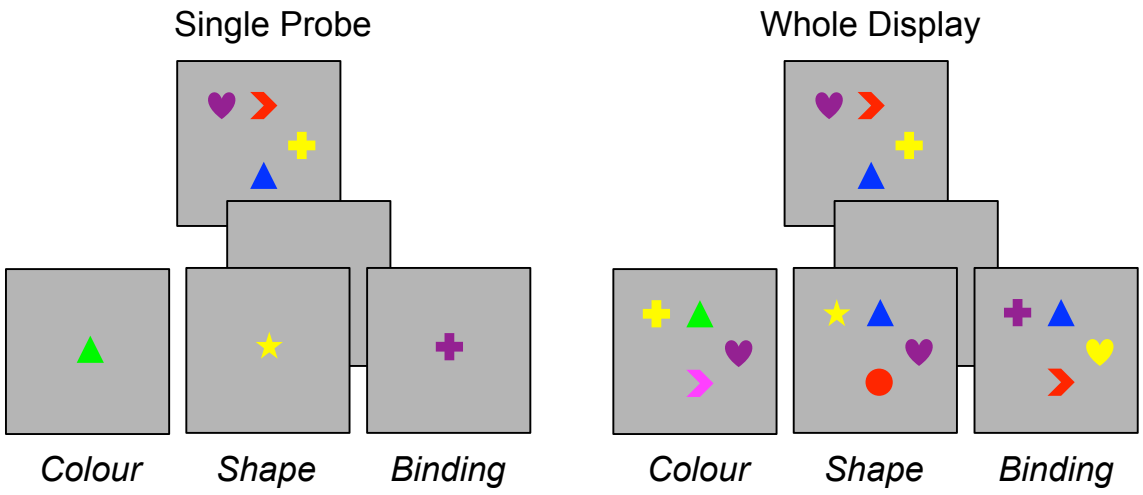

Figure 2.2: Examples of possible memory and test arrays in the two change detection tasks used in Experiment 1. The left panel shows the single probe task and the right the whole display task. Items are not drawn to scale.

the recombination of a colour and shape that were both in the initial set but not paired in the same object.

For the whole display task, participants were instructed that half of the time the objects in the test array would be identical to those in the memory array, albeit shuffled around, and the other half of the trials would contain two changes. In the individual feature conditions of this task participants were informed that a change would constitute the introduction of two brand new features to the test objects, with the task-irrelevant features remaining unchanged. In the binding condition a change involved two shapes swapping colours, thereby creating two new combinations of colour and shape.

Half of our participants completed the single probe task first and the other half the whole display task first. The order of memory conditions was counterbalanced with the constraint that each participant completed the different memory conditions in the same order for each version of the task. Participants were given detailed instructions prior to the experiment with visualisations of change and no-change trials for each of the 6 block types. During the experiment participants were given written instructions informing them of the nature of the block (what probe type and memory condition) along with a verbal description of the type of change they were supposed to look out for. Each block consisted of 8 practice trials and 72 experimental trials.

Half of the trials in each block were change trials and half contained no-change, evenly distributed across the two set sizes (4 and 6). Participants responded to the test array by pressing keys labelled *'SAME'* or *'DIFF'* which corresponded to the *'z'* and *'m'* keys on the keyboard, respectively.

Analysis

Standard inferential statistics based on null hypothesis significance testing (NHST) are fraught with problems (Wagenmakers, 2007). By considering a single hypothesis (the null hypothesis) p-values overstate the evidence against the null and thus in favour of the unspecified 'alternative' (Berger & Sellke, 1987; Sellke, Bayarri, & Berger, 2001). This, amongst other things, has led to the suggestion that researchers should favour *estimation* approaches centered around confidence intervals (CIs; the so called 'new statistics', Cumming, 2013). However, given that CIs are intrinsically related to p-values (Kruschke, 2013) these new statistics do not appear to be able to shake their epistemic problems (Lee, 2014). Further, in the absence of a complete power analysis, failure to reject the null hypothesis leads to an uncertain state where the experimenter is unable to argue for the absence of an effect. In order to provide evidence for the null hypothesis it is necessary to adopt a model comparison approach.

Bayesian analysis offers an alternative to standard approaches to both estimation and hypothesis testing that allows the interpretations usually erroneously bestowed on p-values and CIs (see, Gigerenzer, 2004; Hoekstra, Morey, Rouder, & Wagenmakers, 2014). Here we adopt Bayesian approaches to both estimation and model comparison. There is currently a debate as to which approach provides more principled inference (see, e.g., Kruschke, 2011; Rouder, Morey, et al., submitted) but as we outline in more detail below, each approach comes with its own benefits for specific questions. We now outline the general analysis approach used throughout the thesis.

Estimation As noted in the Introduction section (Chapter 1) analysis of raw accuracy data presents a number of problems and when standard normal models

are applied (e.g. ANOVA) erroneous conclusions can become commonplace (Dixon, 2008; Jaeger, 2008). This is a particular problem when assessing age-differences in accuracy given that older adults tend to be less accurate and consequently more variable. A principled approach to assessing experimental effects on accuracy is to use a *generalised linear model*; here we opt for a hierarchical logit model. This approach is outlined in greater detail in Appendix A but can be summaraised as follows.

The log odds of a correct response on a given trial is modelled as a linear combination of three components:

1. A grand mean parameter reflecting average overall accuracy (β_0).

2. Deflections from the grand mean reflecting main- and interaction-effects of group or experimental factors ($\boldsymbol{\beta}$). These deflections estimate the change in log odds accuracy, relative to the grand mean, associated with being in a specific condition (main effect) or combination of conditions (interaction) and can be used to construct specific contrasts to test hypotheses (Kruschke, 2015). As described by Ntzoufras (2009) they are constrained to sum-to-zero via the use of effects coded variables (see Appendix A for more detail).

3. Finally, as is common in the analysis of repeated measures designs with observations clustered within individuals, we include a random effect of participant with a mean of zero and standard deviation estimated from the data (σ_s) (Gelman & Hill, 2007).

Prior distributions on these parameters were selected to be mildly informative, thus allowing the data to guide our inference (Gelman, Jakulin, Pittau, & Su, 2008; Kruschke, 2015). For each analysis 50000 samples were taken from the joint posterior distribution across 4 independent MCMC chains using JAGS (Just Another Gibbs Sampler, Plummer et al., 2003) after a burn-in period of 5000 samples. All reported chains had converged on a stable distribution as indicated by a multivariate BGR statistic of ≈ 1 (Brooks & Gelman, 1998). These MCMC chains were not thinned (Link & Eaton, 2012) and wherever possible we ensure that for crucial parameters

the *effective sample size* (ESS, Kass, Carlin, Gelman, & Neal, 1998)—the number of independent samples accounting for autocorrelation—is at least 10000 (as per the recommendations of Kruschke, 2015).

For these analyses the crucial quantities of interest are the deflection parameters (component 2 above) which reflect main- and interaction-effects, however for completeness tables are presented for each analysis with posterior summaries of all parameters. With these deflection parameters we can construct specific hypotheses tests, or contrasts, using the general approach outlined by Kruschke (2015). These contrasts are reported in the text along with their 95% highest density intervals (HDIs), which reflect parameters values that are credible given the data (Kruschke, 2015). A contrast, for example between two conditions, that is credibly non-zero (i.e. the HDIs exclude zero) is taken as evidence for a difference, however consideration is always given to the *magnitude* of the effect. A primer on interpreting the size of effects on the log odds scale is given in Appendix A.

Model Comparison In change detection research it is common to summarise performance using a measure that attempts to separate the observer's sensitivity from response bias (see Chapter 8 for more detail) and the equations (2.1, 2.2, and 2.3) used here are an instance of this practice. Typically main effects and interactions in the resulting estimates are assessed with NHSTs, such as ANOVA. However, as detailed above NHST does not allow one to state evidence in favour of the null hypothesis and thus ends up being biased against it. This is a particular problem when assessing specific age-related deficits as failure to reject an age-group by condition interaction does not qualify as evidence against such an interaction. Bayes factors offer an intuitive, and increasingly popular, method of stating evidence for or against effects of interest (Edwards, Lindman, & Savage, 1963). Given two hypotheses, for example an interaction between group and condition versus no such interaction, Bayes factors summarise the ratio by which the prior odds of these hypotheses should be modified to obtain the posterior odds (given the data). Thus they provide a good summary of the *weight of evidence* conferred by the data for competing accounts (Kass & Raftery, 1995).

Here we use the default Bayes factors of Rouder, Morey, Speckman, and Province (2012) as implemented in the `BayesFactor` package in `R` (R. D. Morey & Rouder, 2015; R Core Team, 2015). The `BayesFactor` package uses the default Jeffreys-Zellner-Siow (JZS) family of priors outlined by Rouder et al. (2012) in which Cauchy distributions are placed on *effect size* rather than the raw scale of measurement. The analyst is given control over the Cauchy scale parameter—and thus the prior density given to effects within a certain range—and we used the default setting of 0.5. Different models are defined by the presence or absence of priors on main or interaction effects. The marginal likelihoods of models given the data are then compared to yield a Bayes factor for one model relative to a competitor.

When the number of effects is small (< 3) we adopt the approach outlined in the tutorial of Rouder, Morey, Verhagen, Swagman, and Wagenmakers (in press) which is to only consider models in which interactions are accompanied by corresponding main effects; this greatly reduces the number of to-be-considered models. The full output of `BayesFactor` analyses are presented in Tables allowing the reader to compare any of the constructed models they wish, while focus is devoted to a handful of key model comparisons. Further using the 'winning model' from our default Bayes factor analysis we can obtain posterior samples of model parameters in order to estimate effects of interest[4]. When the number of effects in the design is large (≥ 4) the number of possible models becomes unwieldy. In this case a full model containing all main effects and interactions is compared to reduced models omitting a single component at a time.

Results

Proportion correct

The proportion of correct responses across each task, condition, and set size is shown in Figure 2.3. As previously described, we fit a hierarchical logit model using JAGS

[4]One may wonder why we have elected to analyse accuracy and estimates of items in VWM in different ways. Currently the default Bayes factors of Rouder et al. (2012) assume a continuous outcome variable making them appropriate for our number of items metric. However, for correct/ incorrect data it would be inappropriate to use a normal model (see, Dixon, 2008; Jaeger, 2008, and the Introduction section for more discussion) thus we use a logit model.

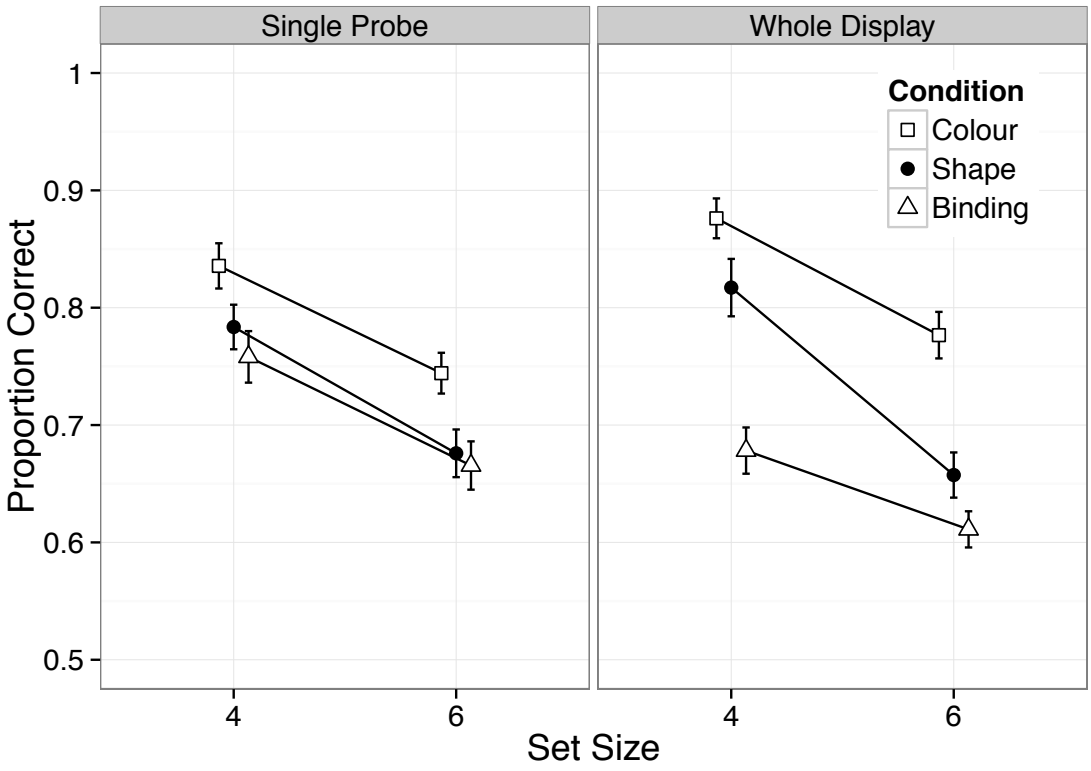

Figure 2.3: Proportion correct for Experiment 1. Error bars are ± standard error.

with the log odds of a correct response modelled as a linear function of effects coded main and interaction effects (deflections from the grand mean). Table 2.1 presents posterior quantities of the model parameters along with ESS when accounting for auto-correlation (Kass et al., 1998).

With the posterior samples the analyst can construct contrasts using the sum-to-zero deflection parameters of the model to test specific hypotheses. Unlike NHST, in which multiple assessments of the data increase the likelihood of false discovery, this approach allows as many (or as few) contrasts to be performed, as the single posterior distribution does not change (Kruschke, 2015)—despite this we focus on a few key contrasts.

As is clear from Figure 2.3 participants were more likely to respond correctly at set size 4 compared to set size 6, 0.580 [0.488, 0.673]. The values given in brackets denote the lower and upper limits of the posterior 95% HDI (see above). In the case of set size the difference is clearly non-zero and corresponds to a difference on the probability scale of approximately 0.108 [0.089, 0.126].

Table 2.1: Posterior quantities from logit model for Experiment 1

			95% HDI		
Parameter	Mean	Median	lower	upper	ESS
β_0	1.117	1.116	0.976	1.263	1535.650
$\boldsymbol{\beta}_1$: (1) Shape	-0.050	-0.049	-0.114	0.014	20321.460
$\boldsymbol{\beta}_2$: (2) Binding	-0.341	-0.341	-0.402	-0.277	20591.959
$\boldsymbol{\beta}_3$: (3) Set Size 6	-0.290	-0.290	-0.337	-0.244	27773.374
$\boldsymbol{\beta}_4$: (4) Single Probe	0.000	0.000	-0.047	0.046	28610.906
$\boldsymbol{\beta}_5$: 1×3	-0.066	-0.066	-0.129	0.001	19767.102
$\boldsymbol{\beta}_6$: 2×3	0.099	0.099	0.037	0.163	20863.052
$\boldsymbol{\beta}_7$: 1×4	-0.032	-0.032	-0.096	0.033	21027.556
$\boldsymbol{\beta}_8$: 2×4	0.162	0.162	0.099	0.223	20175.532
$\boldsymbol{\beta}_9$: 3×4	0.024	0.024	-0.022	0.071	28408.687
$\boldsymbol{\beta}_{10}$: $1 \times 3 \times 4$	0.051	0.051	-0.015	0.115	20832.116
$\boldsymbol{\beta}_{11}$: $2 \times 3 \times 4$	-0.065	-0.065	-0.128	-0.003	20789.063
σ_s	0.333	0.326	0.226	0.451	11468.083

Note: The effects coded variables were as follows: (1) Shape = 1, Binding = 0, Colour = -1, (2) Shape = 0, Binding = 1, Colour = -1, (3) SS4 = -1, SS6 = 1, (4) Whole display = -1, Single probe = 1. Interaction contrasts were products of these effects coded variables.

One contrast of particular interest is between performance in the individual feature conditions (colour or shape only) relative to the binding condition. Thus we contrast the average of the colour and shape deflections from the mean to the binding parameter at each step of the MCMC chain. Observers were much more likely to be correct in the individual feature conditions relative to the binding condition, 0.511 [0.416, 0.603]. Contrasting overall performance between the two types of probe we find no clear difference between single probe and whole display, 0.000 [-0.094, 0.092]. These effects are qualified by a clear interaction between condition and probe type, such that the difference between features and binding was less pronounced when using the single probe relative to a whole display, -0.485 [-0.670, -0.297]. Thus we replicate Wheeler and Treisman's original finding that, considering the probability of a correct response, binding is impoverished by a whole display.

Further, there was a tendency for the effect of increasing set size to be somewhat smaller with a single probe relative to a whole display, but this contrast was not credibly different from zero, -0.096 [-0.284, 0.086]. Finally the two way interaction

Table 2.2: Mean (standard error) hit and false alarm rates accross experimental conditions in Experiment 1

		Single Probe		Whole Display	
Condition	Set Size	h	f	h	f
Colour	4	0.90 (0.01)	0.22 (0.03)	0.93 (0.01)	0.18 (0.03)
	6	0.87 (0.02)	0.38 (0.03)	0.73 (0.02)	0.18 (0.03)
Shape	4	0.81 (0.02)	0.25 (0.03)	0.77 (0.03)	0.14 (0.02)
	6	0.74 (0.03)	0.38 (0.03)	0.55 (0.03)	0.23 (0.03)
Binding	4	0.75 (0.02)	0.23 (0.03)	0.59 (0.03)	0.23 (0.03)
	6	0.75 (0.02)	0.42 (0.04)	0.51 (0.02)	0.29 (0.03)

between features versus binding and probe type was modulated by set size; as shown in Figure 2.3 the two methods of probing are more similar at set size 6 relative to 4, -0.391 [-0.770, -0.018]. This seemed to be largely driven by shape only relative to binding (-0.464 [-0.896, -0.033]), as the contrast with colour (-0.318 [-0.783, 0.142]) was not clearly distinguishable from zero, but given the width of the HDIs these complex interaction contrasts should be interpreted with caution. It should be noted that this pattern is not in line with the expectation of binding specific whole display interference, as one would expect a larger set size effect in the binding condition as the number of interfering test stimuli increases.

Estimated number of items in VWM

The rate of hits and false alarms across the different experimental conditions is presented in Table 2.2. Using Equations 2.1, 2.2, and 2.3 we calculated k for each participant across each combination of factors (see Figure 2.4).

Table 2.3 provides the results of the default Bayes factor analysis for Experiment 1. The Bayes factors provided in the table are for the model (made up of priors on main effect/ interactions) relative to a null model which only contains a random participant effect. Providing the results in this manner allows the reader to compare any models they like; here we focus on a selection of key comparisons. When reporting Bayes factors the subscripts given denote the models being compared; thus $B_{1,0}$ provides the evidence for model 1 relative to model 0 (as shown in row 1 of Table 2.3), whereas $B_{0,1}$ denotes the evidence for the null model relative to model 1

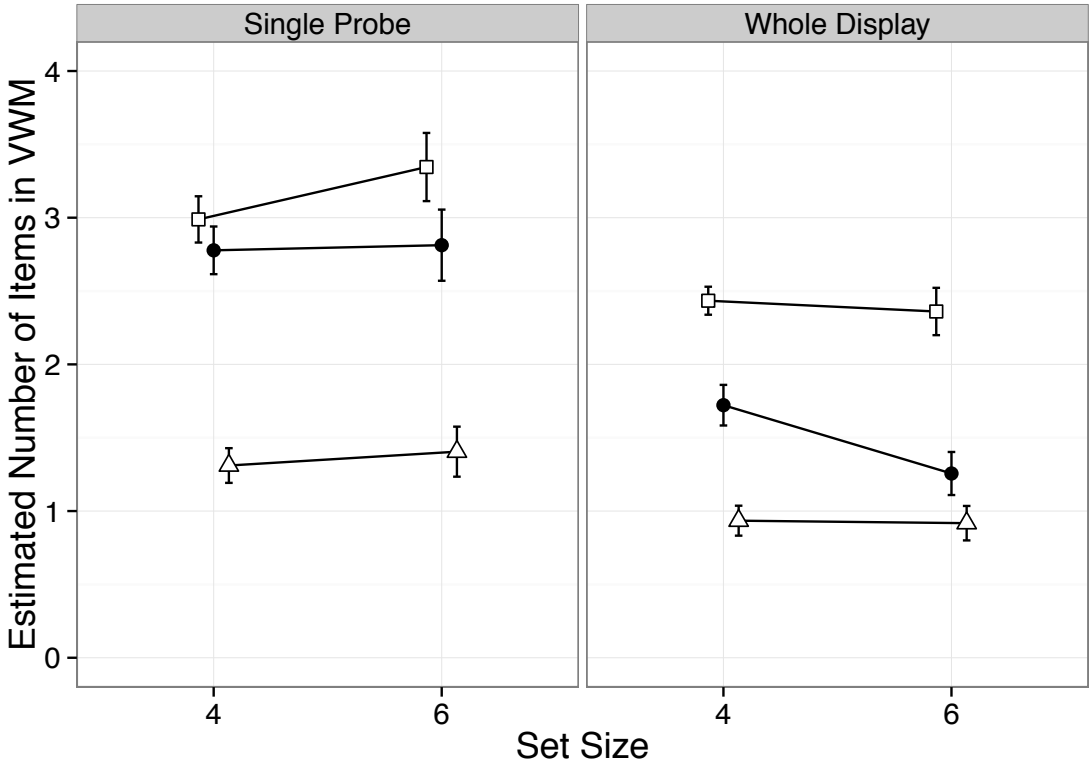

Figure 2.4: Estimated number of items in visual working memory for Experiment 1. Error bars are ± standard error.

(note: $B_{0,1} = 1/B_{1,0}$).

Contrasting models 1 and 7 we can gauge the evidence for the interaction between probe type and memory condition. The model including this interaction is favoured over the model omitting it by around 569-to-1 (that is, $B_{1,7} \approx 569$). The winning model did not contain set size and comparing models 1 and 2 we find that omitting set size is favoured by approximately 8-to-1. However, the comparison regarding the interaction between probe type and set size was not convincing ($B_{3,2} = 0.58$), corresponding to odds of 1.7-to-1 in favour of its omission.

In order to probe these trends further we took 10000 samples from the posterior distribution of the winning model. Specific contrasts revealed that estimates were 0.825 [0.674, 0.986] higher for the single probe task relative to the whole display task. Contrasting individual feature and binding k, as shown in Figure 2.4, there was a clear benefit for features, 1.307 [1.141, 1.468]. This was qualified by the presence of the probe type by condition interaction; the disparity between features

Table 2.3: Log Bayes factors for Experiment 1

	Model	$\log(B_{M,0})$	% error
1	$k \sim$ PT + C + PT:C + ID	117.36	0.78
2	$k \sim$ PT + C + PT:C + SS + PT:SS + ID	115.87	1.33
3	$k \sim$ PT + C + PT:C + SS + ID	115.33	1.07
4	$k \sim$ PT + C + PT:C + SS + PT:SS + C:SS + ID	114.82	1.15
5	$k \sim$ PT + C + PT:C + SS + C:SS + ID	114.23	1.36
6	$k \sim$ PT + C + PT:C + SS + PT:SS + C:SS + PT:C:SS + ID	113.10	2.60
7	$k \sim$ PT + C + ID	111.02	0.37
8	$k \sim$ PT + C + SS + PT:SS + ID	109.37	2.02
9	$k \sim$ PT + C + SS + ID	108.98	0.65
10	$k \sim$ PT + C + SS + PT:SS + C:SS + ID	108.21	3.28
11	$k \sim$ PT + C + SS + C:SS + ID	107.78	1.37
12	$k \sim$ C + ID	69.20	0.22
13	$k \sim$ C + SS + ID	67.17	0.47
14	$k \sim$ C + SS + C:SS + ID	65.54	1.16
15	$k \sim$ PT + ID	21.45	0.38
16	$k \sim$ PT + SS + ID	19.39	0.64
17	$k \sim$ PT + SS + PT:SS + ID	18.69	1.58
18	$k \sim$ SS + ID	-2.05	0.23

Note: All Bayes factors are relative to a null model with a random participant effect only (model 0: $k \sim$ ID). PT = Probe Type, C = Condition, SS = Set Size, ID = participant ID, and ':' denotes an interaction effect

and binding was 0.552 [0.233, 0.872] larger in the single probe relative to the whole display condition. This can be seen clearly in Figure 2.4 where binding appears to be specifically impoverished in the single probe condition.

Discussion

Using slightly different methodology, primarily comparing probe type within participants, we recreate the pattern found by Wheeler and Treisman (2002). The whole display task had a much greater disparity between the feature and binding conditions, in terms of the likelihood of a correct response, relative to the single probe task (see also, Kondo & Saiki, 2012; Yeh et al., 2005). However, as outlined above, comparing the probability of correct response between these tasks is misleading and the single probe task will overestimate binding (Cowan et al., 2013).

Using simple processing models to estimate the number of items in VWM from patterns of hits and false alarms reveals a completely different story. The probe type by condition interaction remains but is a mirror image of that seen for correct responses; the disparity between features and binding is greater in the single probe task. This was highly unexpected, however a potential account lies in our design

relative to the original methods of Wheeler and Treisman.

In the individual feature conditions of Experiment 1 the task-irrelevant feature was presented at test. This is somewhat different to the approach of Wheeler and Treisman (2002) as in their single probe colour-shape experiment (4B) the task irrelevant feature was held constant at test (colours were presented in squares and shapes in black) whereas in the whole display experiment (4A) the irrelevant features were left unchanged. It is possible that participants used the task-irrelevant feature to restrict memory search, via the use of bound representations. For example, knowing that the circle was previously coloured red leads to greater certainty that there is a new feature present when red appears in a square. This information, if used, would be of greater use in the single probe task where search can be restricted to a single VWM item through the use of the task-irrelevant feature, whereas for the whole display it is still the case that the participant has to search for each test item in VWM until a change is found. This may account for greater disparity between features and binding in the single probe task in terms of the estimated number of items in VWM, as feature performance is being boosted by restricted search.

There have been recent investigations into the use of task-irrelevant location information in the change detection task. Both accuracy and latency data suggest that participants do not make use of location information in the single probe task to restrict memory search and instead appear to perform an exhaustive search of VWM for the probed feature (Z. Chen & Cowan, 2013; Cowan et al., 2013; Gilchrist & Cowan, 2014). Nevertheless, we assessed whether we could recreate our initial findings when task-irrelevant features were held constant in probe arrays testing VWM for individual features.

2.3 Experiment 2 – Removing Irrelevant Features

In Experiment 1 we found that, in terms of the estimated number of items in VWM, the disparity between features and bindings was greater for the single probe task

relative to the whole display task. As outlined above, it is possible that participants used irrelevant information to guide change detection decisions and that this information may be of greater use in the single probe task; thus in Experiment 2 we removed the irrelevant feature information at test.

Methods

Participants

Twenty-four new participants (aged 18–30, 13 females) were recruited from the student community of the University of Edinburgh. Each participant received payment of £ 5 for the 45 minute testing session.

Stimuli, Design, and Procedure

The stimuli, design, and procedure used in Experiment 2 were identical to those of Experiment 1 with the only difference being that in the individual feature conditions the task-irrelevant feature was held constant at test. In the colour condition test items were presented in squares (a shape outside of the task set) and in the shape condition items were filled in black (a colour outside of the task set).

Results

Proportion correct

Figure 2.5 depicts proportion correct for Experiment 2. Visually there is little to distinguish the pattern of performance from Experiment 1, and the results of our modelling largely corroborate this.

Table 2.4 presents quantities from the posterior distribution of Experiment 2. Unsurprisingly, correct responses were more likely when 4 items were to-be-remembered relative to 6, 0.486 [0.393, 0.577]. And once again correct responses were more likely in conditions assessing feature memory relative to binding, 0.431 [0.337, 0.523].

However, unlike the previous study, there appeared to be an overall effect of probe type; Correct responses were less probable with a single probe relative to a

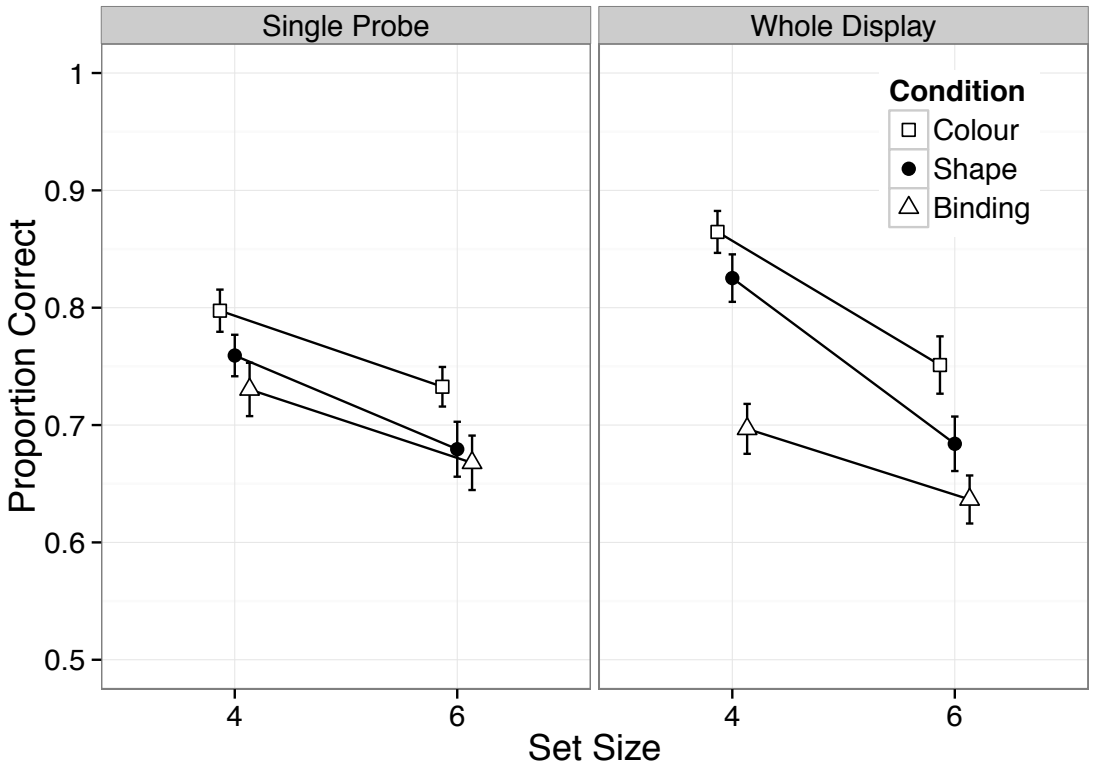

Figure 2.5: Proportion correct for Experiment 2. Error bars are ± standard error.

whole display, -0.119 [-0.211, -0.030]. While zero falls outside of the HDI this is clearly a very small effect and corresponds to a difference of approximately -0.023 [-0.040, -0.006] on the probability scale. Thus for practical purposes the overall effect of probe type on the probability of a correct response appears to be fairly negligible.

Crucially, in line with Experiment 1 and Wheeler and Treisman (2002), the difference between feature and binding performance was less pronounced with a single probe relative to whole display, -0.411 [-0.597, -0.222]. There was also an interaction between set size and probe type such that the effect of increasing the number of to-be-remembered items was less pronounced with a single probe, -0.253 [-0.434, -0.074]. Again the two-way interaction between condition and probe type was modulated by set size as the whole display 'interference' effect was less pronounced at set size 6, -0.420 [-0.788, -0.041]. The magnitude of this effect appeared to be quite similar for the contrast between colour and binding (-0.422 [-0.874, 0.021]) and between shape and binding (-0.418 [-0.842, 0.011]). Again this is not in line with binding specific interference, which would predict a larger effect of set size in the

Table 2.4: Posterior quantities from logit model for Experiment 2

			95% HDI		
Parameter	Mean	Median	lower	upper	ESS
β_0	1.080	1.078	0.935	1.230	1500.473
$\boldsymbol{\beta}_1$: (1) Shape	0.002	0.002	-0.061	0.066	19688.410
$\boldsymbol{\beta}_2$: (2) Binding	-0.288	-0.288	-0.348	-0.225	19791.540
$\boldsymbol{\beta}_3$: (3) Set Size 6	-0.243	-0.243	-0.288	-0.197	27891.260
$\boldsymbol{\beta}_4$: (4) Single Probe	-0.060	-0.060	-0.105	-0.015	28496.956
$\boldsymbol{\beta}_5$: 1×3	-0.058	-0.058	-0.121	0.007	19872.993
$\boldsymbol{\beta}_6$: 2×3	0.097	0.097	0.036	0.159	20184.902
$\boldsymbol{\beta}_7$: 1×4	-0.049	-0.049	-0.114	0.015	19509.445
$\boldsymbol{\beta}_8$: 2×4	0.137	0.137	0.074	0.199	19638.935
$\boldsymbol{\beta}_9$: 3×4	0.063	0.063	0.018	0.108	28959.320
$\boldsymbol{\beta}_{10}$: $1 \times 3 \times 4$	0.035	0.035	-0.029	0.098	19527.026
$\boldsymbol{\beta}_{11}$: $2 \times 3 \times 4$	-0.070	-0.070	-0.131	-0.007	19474.866
σ_s	0.345	0.338	0.232	0.465	12638.405

Note: The effects coded variables were as follows: (1) Shape = 1, Binding = 0, Colour = -1, (2) Shape = 0, Binding = 1, Colour = -1, (3) SS4 = -1, SS6 = 1, (4) Whole display = -1, Single probe = 1. Interaction contrasts were products of these effects coded variables.

binding condition.

Estimated number of items in VWM

Table 2.5: Mean (standard error) hit and false alarm rates accross experimental conditions in Experiment 2

		Single Probe		Whole Display	
Condition	Set Size	h	f	h	f
Colour	4	0.89 (0.01)	0.29 (0.03)	0.87 (0.02)	0.14 (0.02)
	6	0.83 (0.02)	0.37 (0.03)	0.66 (0.04)	0.16 (0.03)
Shape	4	0.74 (0.03)	0.22 (0.02)	0.84 (0.03)	0.19 (0.02)
	6	0.60 (0.04)	0.25 (0.03)	0.54 (0.03)	0.17 (0.02)
Binding	4	0.72 (0.03)	0.26 (0.03)	0.61 (0.03)	0.22 (0.02)
	6	0.72 (0.02)	0.38 (0.04)	0.50 (0.03)	0.22 (0.02)

The hit and false alarm rates observed in Experiment 2 are presented in Table 2.5 and we estimated the number of items in VWM across the experimental factors as before (see Figure 2.6). Table 2.6 shows that, as in the analysis of Experiment 1, the winning model contains main effects of condition and probe type, along with their

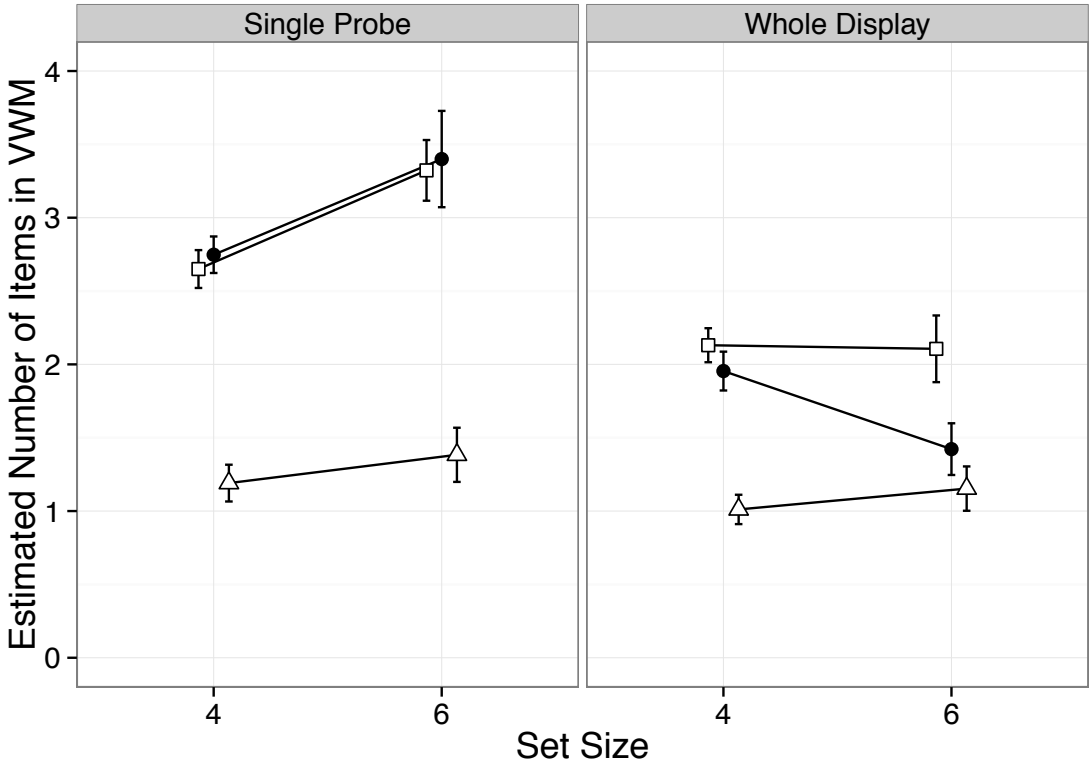

Figure 2.6: Estimated number of items in visual working memory for Experiment 2

interaction. However, in addition set size and the interaction between probe type and set size also appear in this winning model. Comparing models 1 and 7 in Table 2.6 we find that the model including the probe type by condition interaction is favoured over the model omitting it to an overwhelming degree ($B_{1,7} = 1.454186 \times 10^4$). While the weight of evidence is against an overall effect of set size ($B_{18,0} = 0.33$) the evidence for the probe type by set size interaction in this experiment is strong ($B_{1,5} = 62.73$). Finally, by comparing models 2 and 3 we can gauge the evidence for the three-way interaction. This comparison favours the interaction model ($B_{2,3} = 1.73$), but only marginally, therefore we prefer the more parsimonious model 1 ($B_{1,2} = 4.69$).

To probe these trends further we again took 10000 samples from the posterior distribution of the winning model. As with the previous analysis, estimates of the number of items in VWM were greater in the single probe task relative to whole display, 0.805 [0.629, 0.983] and features were more readily stored relative to bindings, 1.263 [1.073, 1.448]. The interaction between probe type and condition

Table 2.6: Log Bayes factors for Experiment 2

	Model	$\log(B_{M,0})$	% error
1	$k \sim$ PT + C + PT:C + SS + PT:SS + ID	88.24	1.12
2	$k \sim$ PT + C + PT:C + SS + PT:SS + C:SS + PT:C:SS + ID	86.70	2.07
3	$k \sim$ PT + C + PT:C + SS + PT:SS + C:SS + ID	86.15	1.57
4	$k \sim$ PT + C + PT:C + ID	84.26	0.82
5	$k \sim$ PT + C + PT:C + SS + ID	84.10	1.56
6	$k \sim$ PT + C + PT:C + SS + C:SS + ID	82.02	0.78
7	$k \sim$ PT + C + SS + PT:SS + ID	78.66	0.61
8	$k \sim$ PT + C + SS + PT:SS + C:SS + ID	76.53	1.98
9	$k \sim$ PT + C + ID	75.34	0.78
10	$k \sim$ PT + C + SS + ID	75.03	1.13
11	$k \sim$ PT + C + SS + C:SS + ID	72.88	0.96
12	$k \sim$ C + ID	46.36	0.32
13	$k \sim$ C + SS + ID	45.69	0.33
14	$k \sim$ C + SS + C:SS + ID	43.42	0.58
15	$k \sim$ PT + SS + PT:SS + ID	19.10	1.65
16	$k \sim$ PT + ID	18.46	0.29
17	$k \sim$ PT + SS + ID	17.51	0.42
18	$k \sim$ SS + ID	-1.11	0.49

Note: All Bayes factors are relative to a null model with a random participant effect only (model 0: $k \sim$ ID). PT = Probe Type, C = Condition, SS = Set Size, ID = participant ID, and ':' denotes an interaction effect

was again born out of a greater disparity between features and bindings in the single probe task, 0.851 [0.483, 1.221]. It is worth noting that the magnitude of this interaction effect is greater than the corresponding value for Experiment 1 (0.552 [0.233, 0.872]), although the two HDIs show a fair amount of overlap.

In line with the findings from comparing models the difference between estimates at set size 4 and 6 was small and not convincingly different from zero (-0.176 [-0.354, -0.007]). Finally, the effect of increasing the number of to-be-remembered items was clearly less pronounced in the whole display task relative to single probe (-0.604 [-0.945, -0.24]). As shown in Figure 2.6, estimates are larger at set size 6 relative to 4 in the single probe task whereas there is comparatively little effect of set size with a whole display. Of course the plot reveals a slightly more complicated picture with a negative effect of set size in the whole display shape condition, however, as mentioned above we prefer this simpler account given the weak evidence for the three-way interaction

Discussion

In Experiment 2 we removed the irrelevant feature information from the test displays in our colour and shape conditions as this information may have been used to restrict memory search in Experiment 1. The pattern of proportion correct (Figure 2.5) was largely the same as Experiment 1 and the findings of Wheeler and Treisman in that the whole display task appeared to specifically impair binding performance.

Examining the estimated number of items in VWM revealed that, as in Experiment 1, the disparity between feature and binding k was greater in the single probe task. Removing the irrelevant feature dimension from the individual feature condition test displays did not greatly alter this aspect of the results, if anything the effect was magnified. Therefore it appears that the presence of additional, potentially useful, information has little effect on the disparity between the number of features and bindings observers can retain and use to perform change detection. In outlining the rationale for the third experiment we discuss another potential explanation for the unexpected probe type by condition interaction clearly visible in Figures 2.4 and 2.6.

There were some departures from the original results of Experiment 1, namely the appearance of an interaction between probe type and set size. The effect of increasing the number of to-be-remembered items was greater in the single probe task (larger estimates at set size 6) relative to the whole display task (slight negative effect). However, from the above separate analyses of Experiments 1 and 2 it is difficult to conclude that there are clear differences between the experiments. Therefore we ran an extra Bayes factor analysis with an additional factor of experiment. Adopting the approach we did in the above analysis would result in a large number of models (167) therefore we took a different approach in which a 'full' model with all main effects and interactions was compared to models omitting a single component (as implemented by specifying *'top'* in the `BayesFactor` package). This greatly reduces the number of models (to 16) and while it is more restrictive than the above approach it is sufficient for the present purposes.

Comparing the full model (model F) to a reduced one (model R) omitting ex-

periment we find that the reduced model is preferred by approximately 5-to-1. So there is no suggestion that overall estimates of k differed between Experiments 1 and 2. The weight of evidence is also against a modulatory effect of Experiment on probe type (reduced model favoured by 8-to-1) and set size (2-to-1), although in the latter case the evidence is rather weak. Comparing Figures 2.4 and 2.6 there is no clear, systematic difference between the experiments in terms of the overall effect of set size. There is, however, evidence that the effect of condition was not the same across experiments as the Bayes factor for the model omitting the condition by experiment interaction relative to the full model was 0.2 implying around 5-to-1 support for the inclusion of this interaction, we return to this shortly.

As for higher order interactions including experiment the evidence always favoured the reduced models omitting the interaction of interest. In the case of the three way interactions including condition and probe type ($B_{R,F} = 7.47$) and including condition and set size ($B_{R,F} = 4.65$) the evidence against the interactions was fairly strong. However, the evidence was less clear for the three way interaction including probe type and set size ($B_{R,F} = 2.77$). While this does not support the idea that the probe type by set size interaction differed between experiments the small Bayes factor is in line with the appearance of this interaction in the winning model of Experiment 2. Finally, there was no indication of the 4 way interaction of all experimental factors by experiment ($B_{R,F} = 4.65$).

Returning to the condition by experiment interaction which gained support in the above analysis. When we assess the marginal mean estimates across the two experiments we find that colour k was higher in Experiment 1 ($M = 2.78$) with irrelevant shape present than Experiment 2 ($M = 2.55$) when this information was lacking. On the other hand shape shows the opposite pattern with higher estimates in Experiment 2 ($M = 2.38$) without colour present relative to Experiment 1 ($M = 2.14$). It is difficult to make any strong argument for the role of irrelevant features in VWM with effects in the order of 0.2 of an item. However others have previously suggested that colour and shape may be 'asymmetrically bound'. Ecker, Maybery, and Zimmer (2013) found that task-irrelevant variation in object colour

between study and test interfered with shape change detection but variation in shape did not have the same effect on colour change detection. They suggest that colour is obligatorily bound to shape but not vice versa. They suggest that this asymmetric binding may have occurred because colour is a particularly salient feature and thus detecting colour changes can be done without the aid of other features. However, our pattern of results suggests that the presence of shape benefits colour change detection, whereas the presence of colour hinders shape detection. The reasons for this are unclear, however given the size of the differences, the practical consequences of including irrelevant features or holding them constant at test appear to be minimal.

One common finding in Experiments 1 and 2 was reliably lower estimates of k in the whole display task, regardless of the memory condition. This might suggest that processing multiple display items at test is damaging to VWM maintenance in general, and not specifically to the maintenance of feature bindings. It may be argued that this whole display interference effect reflects interference specifically with 'fragile visual short-term memory', a store distinct from iconic memory but far more fragile than VWM (Sligte, Scholte, & Lamme, 2008). This seems possible given that fragile memory is particularly susceptible to overwriting from objects sharing both location and features with the memoranda (Pinto, Sligte, Shapiro, & Lamme, 2013). Fragile memory may be better utilised to support task performance with the single probe as this is presented in a central position, not previously occupied by memory items. However, given the proposed importance of feature overlap, this account would presumably predict a larger discrepancy between the two probe types in Experiment 1 where the test objects shared more features with the initial study items, which we did not find (see above). As we describe in the rationale for the third experiment there are reasons to suspect that this whole display interference effect and the disproportionate binding deficit in the single probe condition are, in fact, artifacts of the modelling approach.

2.4 Experiment 3 – Dual Probe versus Whole Display

In Experiments 1 and 2 we used processing models derived from a slots account of VWM to compare the single probe and whole display tasks. Assessing the estimated number of items in VWM suggested a general whole display interference effect and specifically low estimates in the single probe binding condition. However, it is possible that these patterns of results arose given the necessity to use processing models that imply different maximum possible values for k. For example, in the whole display tasks one of two items is sufficient to detect a change, thus the maximum obtainable value of k is $N - 1$ (this is also the case for single probe binding), whereas for the single probe individual feature task the maximum possible value of k is limited to N.

Averaging over participants to obtain a mean estimate for each condition can bias estimates downwards if some observer's true capacity exceeds the array size (see, Rouder et al., 2011, for more detail on *problematic averaging*). Further, if lapses of attention are present the estimate of k is also biased downwards. Both of these downward biases would have a greater effect in the conditions with the lowest maximum possible capacity estimates and would thus serve to exacerbate any differences between the single probe individual feature conditions and the single probe binding condition. This may help explain the probe type by condition interaction observed in the first two experiments. Further, this downward bias would also serve to reduce estimates from the whole display task relative to the single probe feature conditions, thus the more general whole display interference effect we have been observing may also be a consequence of the different processing models used in calculating k.

Consequently in the third experiment we attempted to physically match the two tasks as closely as possible in order to better compare them without the need for post-hoc modelling. As the whole display task always requires two changes to occur (given the two necessary changes in the binding condition that constitute a feature swap) it seems reasonable to present two items in the 'single' probe task, with

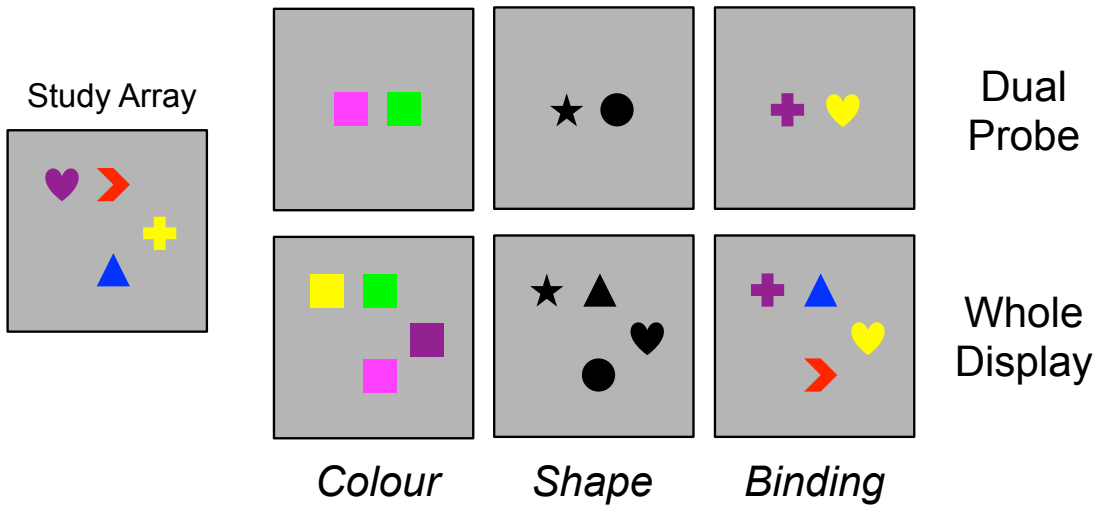

Figure 2.7: Comparison of the test arrays in the dual probe and whole display tasks. Only change trials are shown.

the possibility that neither or both of the objects have changed. This can still be considered in essence a single probe task, as the observer has to only make a single judgement (*same* or *different*), however for clarity we will refer to it as the 'dual probe' task. Figure 2.7 demonstrates this dual probe task for individual features and bindings in comparison to the whole display task. This dual probe versus whole display comparison is more appropriate given that, in both tasks and all conditions, the aim is to detect two changes. Therefore, any inherent difficulties in performing the whole display task, and making several judgments at test, relative to making a single judgement at test should be apparent in the probability of giving a correct response.

While the primary purpose of this third experiment was to physically match the tasks as much as possible rather than using post-hoc modelling it is instructive to outline the principled processing model for this change detection task. For the dual probe the appropriate model is the same in all conditions (feature or binding). If no-change has occurred the observer detects this if either one or both of the unchanged items is in VWM. This is given by c as defined above; therefore the likelihood that the observer makes a false alarm is given by,

$$f = (1 - c)g.$$

When the two changes have occurred (i.e. the two probe items were not in the

original set) and the number of items in VWM is exceeded by the array size an incorrect *same* response can only arise due to guessing. Consequently,

$$1 - h = 1 - g,$$

so long as the array size is sufficiently large. It is clear that the appropriate processing model for the dual probe task is essentially a mirror image of the two-change whole display model, with hits and false alarms switching roles. Combining the above equations we can solve for an estimate of the number of items in VWM,

$$\hat{k} = \frac{2N\hat{h} - \sqrt{\hat{h}}\sqrt{4N(N\hat{f} - \hat{f}) + \hat{h}}}{2\hat{h}}, \tag{2.4}$$

which will give an estimate of k in the dual probe task provided that $k \leq N$, $\hat{h} \geq \hat{f}$, and $\hat{h} > 0$.

Methods

Participants

Twenty-four new participants (aged 18–35, 15 females) were recruited from the student community of the University of Edinburgh. Each participant received payment of £ 5 for the 45 minute testing session.

Stimuli, Design, and Procedure

The stimuli and general procedure used in this experiment was almost identical to that used in Experiment 2. Probe displays in conditions assessing VWM for individual features were presented holding the task-irrelevant feature constant. The only difference was that the single probe task was replaced by a dual probe task in which two items were presented at test (Figure 2.7). Probe items were presented to the left and right of the centre of the screen and did not overlap with any of the initially present memory items. In this task both tested items were either members of the initial memory set or both were different. This was made clear to participants and visualisations of the different changes they were to look out for were given

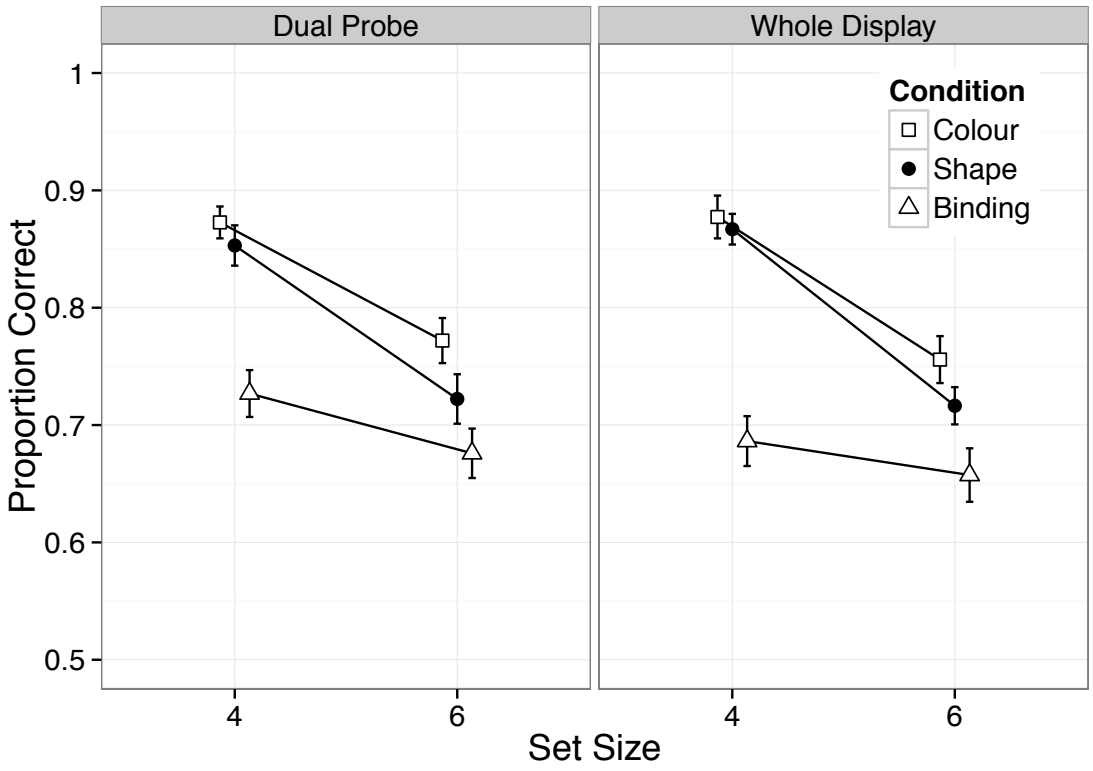

Figure 2.8: Proportion correct for Experiment 3. Error bars are ± standard error.

before starting each block of trials. Requiring the detection of two changes in all conditions of both tasks provides a better method of comparing a single decision at test to multiple decisions.

Results

Proportion Correct

Matching the two tasks in terms of the number of changes that must be detected had a clear effect on the pattern of results. There are no clear, standout differences between the two panels of Figure 2.8, unlike the data from our previous two experiments (see Figures 2.3 and 2.5 for comparison).

The results of our logit model estimation are presented in Table 2.7. As in our previous experiments the likelihood of a correct response was greater in feature conditions relative to the binding condition, 0.692 [0.595, 0.788], and when 4 items were presented relative to 6, 0.620 [0.521, 0.715]. Further, there was no clear difference

Table 2.7: Posterior quantities from logit model for Experiment 3

			95% HDI		
Parameter	Mean	Median	lower	upper	ESS
β_0	1.266	1.265	1.133	1.403	2013.469
$\boldsymbol{\beta}_1$: (1) Shape	0.138	0.138	0.070	0.210	18953.689
$\boldsymbol{\beta}_2$: (2) Binding	-0.461	-0.461	-0.525	-0.397	19617.481
$\boldsymbol{\beta}_3$: (3) Set Size 6	-0.310	-0.310	-0.358	-0.261	26779.204
$\boldsymbol{\beta}_4$: (4) Single Probe	0.020	0.020	-0.028	0.068	26005.156
$\boldsymbol{\beta}_5$: 1×3	-0.134	-0.133	-0.204	-0.064	19224.316
$\boldsymbol{\beta}_6$: 2×3	0.215	0.215	0.151	0.279	18873.958
$\boldsymbol{\beta}_7$: 1×4	-0.042	-0.042	-0.113	0.026	19227.554
$\boldsymbol{\beta}_8$: 2×4	0.051	0.051	-0.012	0.114	18723.056
$\boldsymbol{\beta}_9$: 3×4	0.014	0.014	-0.034	0.061	25250.134
$\boldsymbol{\beta}_{10}$: $1 \times 3 \times 4$	0.023	0.023	-0.047	0.091	19492.050
$\boldsymbol{\beta}_{11}$: $2 \times 3 \times 4$	-0.042	-0.042	-0.108	0.019	19628.023
σ_s	0.312	0.306	0.207	0.431	11632.833

Note: The effects coded variables were as follows: (1) Shape = 1, Binding = 0, Colour = -1, (2) Shape = 0, Binding = 1, Colour = -1, (3) SS4 = -1, SS6 = 1, (4) Whole display = -1, Single probe = 1. Interaction contrasts were products of these effects coded variables.

between the dual probe and whole display testing methods, 0.040 [-0.056, 0.137].

Turning to specific interaction contrasts, with the modification to the paradigm the method of probing did not clearly affect the disparity between feature and binding performance, -0.152 [-0.343, 0.037], especially when compared with the corresponding contrasts from Experiments 1 (-0.485 [-0.670, -0.297]) and 2 (-0.411 [-0.597, -0.222]). Also, as shown in Figure 2.8, the effect of increasing set size from 4 to 6 was not greatly modulated by probe type, -0.055 [-0.244, 0.137]. Finally, unlike the previous experiments, there was no clear interaction between probe type, set size, and the disparity between feature and binding performance, -0.254 [-0.647, 0.114], although the limits of the HDI span a wide range of credible values.

Estimated number of items in VWM

We estimated the number of items in VWM using Equations 2.4 and 2.3 as appropriate (see Figure 2.9). Table 2.9 shows that the winning model from our Bayes factor analysis included probe type, condition, and their interaction. However, comparing

Table 2.8: Mean (standard error) hit and false alarm rates accross experimental conditions in Experiment 3

		Single Probe		Whole Display	
Condition	Set Size	h	f	h	f
Colour	4	0.95 (0.01)	0.21 (0.03)	0.87 (0.03)	0.12 (0.02)
	6	0.85 (0.02)	0.31 (0.03)	0.68 (0.03)	0.17 (0.02)
Shape	4	0.86 (0.02)	0.16 (0.02)	0.84 (0.02)	0.11 (0.02)
	6	0.71 (0.03)	0.27 (0.03)	0.60 (0.03)	0.16 (0.02)
Binding	4	0.68 (0.03)	0.22 (0.03)	0.57 (0.03)	0.19 (0.02)
	6	0.71 (0.03)	0.36 (0.03)	0.55 (0.03)	0.23 (0.03)

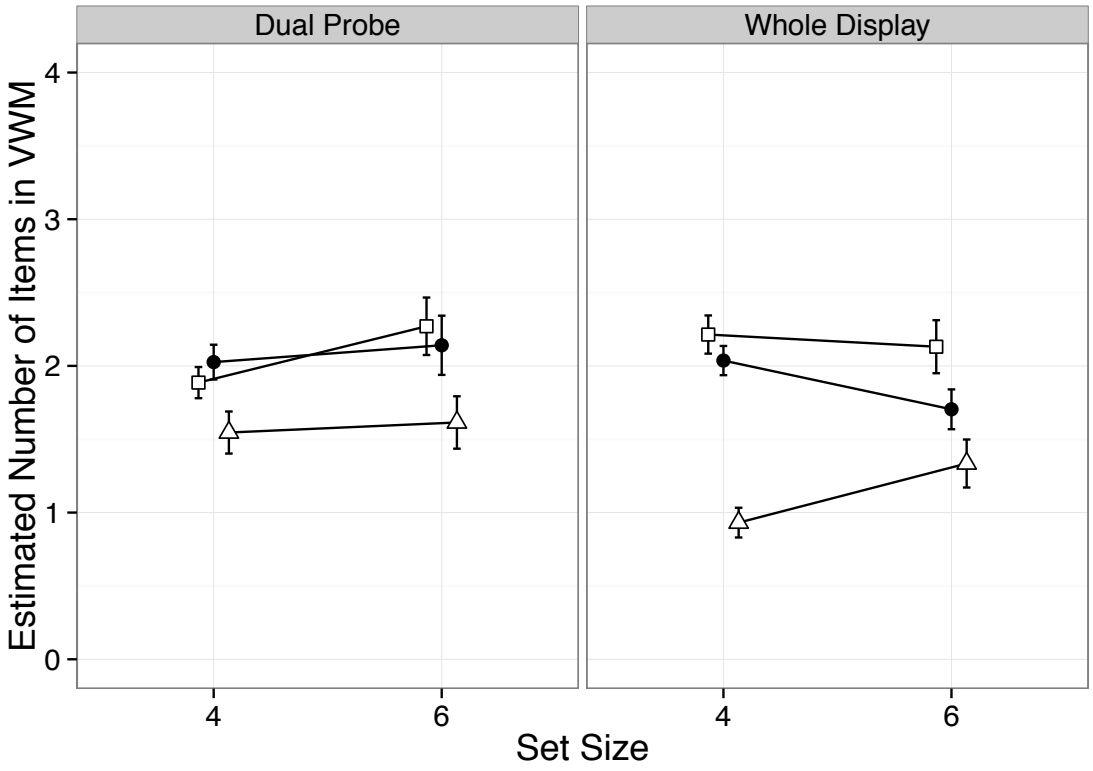

Figure 2.9: Estimated number of items in visual working memory for Experiment 3. Error bars are ± standard error.

models 1 and 2 we find that the evidence for the two-way interaction is fairly weak ($B_{1,2} = 1.98$), so the presence (or, indeed, absence) of such an interaction in the data should be approached with caution.

To gauge the evidence against the set size by condition interaction we compare the best model including set size (model 4) to the corresponding model including the interaction (model 7). Doing this we find that the simpler model is preferred by

Table 2.9: Log Bayes factors for Experiment 3

	Model	$\log(B_{M,0})$	% error
1	$k \sim$ PT + C + PT:C + ID	27.00	0.60
2	$k \sim$ PT + C + ID	26.32	0.66
3	$k \sim$ C + ID	25.71	0.15
4	$k \sim$ PT + C + PT:C + SS + ID	25.59	0.95
5	$k \sim$ PT + C + SS + ID	24.92	0.80
6	$k \sim$ PT + C + PT:C + SS + PT:SS + ID	24.52	1.29
7	$k \sim$ PT + C + PT:C + SS + C:SS + ID	24.40	1.24
8	$k \sim$ C + SS + ID	24.29	0.33
9	$k \sim$ PT + C + SS + PT:SS + ID	23.86	0.94
10	$k \sim$ PT + C + SS + C:SS + ID	23.67	0.68
11	$k \sim$ PT + C + PT:C + SS + PT:SS + C:SS + PT:C:SS + ID	23.48	2.41
12	$k \sim$ PT + C + PT:C + SS + PT:SS + C:SS + ID	23.37	1.18
13	$k \sim$ C + SS + C:SS + ID	23.02	0.41
14	$k \sim$ PT + C + SS + PT:SS + C:SS + ID	22.63	0.92
15	$k \sim$ PT + ID	0.08	0.76
16	$k \sim$ PT + SS + ID	-1.46	0.67
17	$k \sim$ SS + ID	-1.55	0.33
18	$k \sim$ PT + SS + PT:SS + ID	-2.66	0.52

Note: All Bayes factors are relative to a null model with a random participant effect only (model 0: $k \sim$ ID). PT = Probe Type, C = Condition, SS = Set Size, ID = participant ID, and ':' denotes an interaction effect

approximately 3-to-1. A similar comparison for the interaction between probe type and set size yields similar support for the null ($B_{4,6} = 2.91$). Finally, the data are unable to adjudicate in the case of the three way interaction with a Bayes factor of 1 ($B_{11,12} = 1.12$).

As in our previous analyses we took 10000 samples from the posterior distribution of the winning models, to probe these trends further. Estimates of k were slightly higher in the dual probe task relative to the whole display, 0.18 [0.029, 0.332]. This may suggest some additional interference from the whole display probe, however given the magnitude of the effect and that the HDI spans very close to zero this suggestion is not greatly supported by the data. In other words, any interference effect in terms of the number of items in VWM appears minimal. In line with our previous findings, contrasting feature and binding k revealed a distinct feature advantage, 0.673 [0.512, 0.847]. Finally, the disparity between feature and binding performance was -0.338 [-0.652, -0.027] smaller in the dual probe task than in the whole display. The width of the HDI around this specific contrast reveals a great deal of uncertainty as to the size of this effect, this question would clearly benefit from additional data.

Returning to the weak evidence regarding the three-way interaction in this experiment. From Figure 2.9 it looks as if at set size 4 in the whole display binding condition estimates are rather low. Given that the whole display task in this experiment was identical to the previous one (see Figure 2.6) it is not clear why this is the case. However, what is clear is that this pattern is not expected under binding specific whole display interference as the disparity should get larger, not smaller, with increasing set size.

Discussion

Experiment 3 was conducted to obviate the need for post-hoc modelling and better match the single and multiple decision tasks. The results were somewhat surprising; having to make multiple decisions on a whole display did not greatly affect the likelihood of a correct response relative to having to make a single decision on a dual probe. Unlike our previous experiments there was no clear probe type by condition interaction in the accuracy data, reinforcing our suggestion that the initial findings of Wheeler and Treisman (2002)—of binding specific whole display interference—were due to poorly matched tasks.

In terms of the estimated number of items in VWM there was slight evidence for a probe type by memory condition interaction in the direction that the disparity between features and binding was greater for the whole display task. However, as discussed above the pattern of data presented in Figure 2.9 does not conform to the expectations of binding specific whole display interference. Further, if it was the case that the whole display task tended to overwrite bound feature representations then, with these better matched tasks, this should have been clear in the raw accuracy data, which it was not.

The results of the previous two experiments appear to conflict with the present analysis. Estimates of the number of items in VWM from Experiments 1 and 2 appeared to suggest a general whole display interference effect with an additional interaction between probe type and condition such that feature capacity was boosted by a single probe. In Experiment 3, using the raw data with better matched tasks, we

find no clear evidence of whole display interference and no interaction. One possible explanation for this disparity is that our initial processing model-based results were affected by *problematic averaging* (see, Rouder et al., 2011). As mentioned when outlining the appropriate models for each of the change detection tasks we used the formulae only return a valid estimate of k provided that $k \leq N$ (or in some cases $k \leq N - 1$), otherwise they will return N (or $N - 1$). It is reasonable to expect that the majority of participants have true capacities below 4 items, however given that there are vast individual differences in this measure some may have exceeded even our 6 item condition (Vogel & Awh, 2008). As discussed by Rouder et al. (2011), if some participant's true capacity is greater than the array size condition averages across the entire sample are biased downwards. There is the additional issue that, as expressed above, our processing models assume that participants pay attention on every trial. However, it seems reasonable that participants will lapse on some small proportion of trials (Rouder et al., 2008) and in our analyses any lapses of attention will have been attributed to lower capacity. These downward biases will be greatest in the conditions with the lower maximum obtainable capacity; in this case the single probe binding and whole display conditions. Therefore, in terms of comparing performance between tasks greater weight should be given to the results of Experiment 3 without the auxiliary assumptions involved in estimating the number of items in VWM.

2.5 General Discussion

We set out to further investigate a methodological issue in the field, namely does the way in which VWM is probed in the change detection task affect performance? Wheeler and Treisman (2002) found that participants were much less accurate at detecting binding changes relative to feature changes when probed with multiple test items (see also, Kondo & Saiki, 2012; Yeh et al., 2005). However, the single probe task that this was compared to has recently been shown to overestimate binding performance (Cowan et al., 2013; H. Zhang et al., 2010). Therefore, we used simple processing models of VWM and slight modifications to the tasks in order to better

compare them.

Experiments 1 and 2 replicated the finding that in terms of accuracy the discrepancy between feature and binding performance was greater in the whole display task. However, estimates of the number of items in VWM suggested a more general binding cost and generally lower estimates in the whole display task relative to a single probe. Further, these processing model based analyses also suggested that the single probe task led to a much greater binding cost. Given some reservations regarding the use of processing models in an attempt to match the tasks post-hoc, Experiment 3 aimed to physically match the tasks as best as possible. Participants were no more accurate when having to make a single decision on two probe items versus many decisions on multiple test items. Further, there was a general binding cost that was not clearly larger for either of the probe types.

Our findings from Experiment 3, that accuracy was unaffected by a whole display probe are at odds with some previous findings in the literature. For example, Makovski, Watson, Koutstaal, and Jiang (2010) compared two methods of probing VWM and found that they led to different estimates of sensitivity. Both tasks started with the presentation of an initial memory array (coloured squares in different locations) followed by a brief delay; this was then followed by either a single probe change detection (referred to as same-different) task in which a probe item was presented in a previously occupied location or a two-alternative forced choice (2AFC) task in which two colours were probed at a specific location and participants had to choose the correct colour. Makovski and colleagues found that estimates of sensitivity (d') were significantly lower in the 2AFC task relative to the same-different task. They suggest that having to attend to and evaluate multiple stimuli causes interference to fragile VWM at test. The precise nature of this interference was left open but subsequent work has suggested a form of 'overwriting' where previously active representations, tied to a specific location, are updated by the presence of new, potentially task relevant, stimuli (R. J. Allen, Castellà, Ueno, Hitch, & Baddeley, 2015; Alvarez & Thompson, 2009; Logie, Brockmole, & Vandenbroucke, 2009; Ueno, Allen, Baddeley, Hitch, & Saito, 2011; Ueno, Mate, Allen, Hitch, & Baddeley,

2011; Fiacconi & Milliken, 2013).

It is possible that attending to and evaluating two test items led to overwriting of the representations held in VWM. However, there are some potential objections to the approach of Makovski et al. (2010) that cast some doubt on their main conclusions. As mentioned above, the authors reported analyses of d' to compare sensitivity between the two paradigms. This was calculated differently for each paradigm, out of necessity, as the standard formula was used for the same-different task whereas an M-alternative forced choice version was used for the 2AFC task. As noted by the authors, the same underlying sensitivity (in terms of d') results in higher accuracy for the 2AFC task relative to the same-different task, as response bias is assumed to play less of a role. Thus, for the same observed accuracy, estimates of d' will be lower in the 2AFC task relative to the single probe.

However, taking a high-threshold approach to these tasks we find that this may unfairly penalise the 2AFC task, artificially lowering estimates of sensitivity. Assuming participants use location information when making the same-different judgement the slots conception predictions for hit and false-alarm rate are given by,

$$h = d + (1 - d)g$$

$$f = (1 - d)g$$

where d is the probability that the probed item is in VWM, $\min(k/N,\ 1)$, and g is the probability of guessing change when the probe is outside VWM (Cowan, 2001; Rouder et al., 2011).

For the 2AFC it is not possible to separate hits and false-alarms as a target is present on every trial. Therefore, the probability that a correct answer is given in this task is, $p(c)^{\text{2AFC}} = d + (1 - d)\frac{1}{2}$, where d is defined as above and observers have a 50% chance of guessing correctly when the probe is not in VWM.

We can directly compare the above processing models by expressing the single probe model in terms of probability correct. With equal change/no-change probabilities $p(c)^{\text{SP}} = \frac{1}{2}(h + [1 - f])$, so we can combine the predictions and simplify for,

$$p(c)^{\text{SP}} = \frac{k}{2N} + \frac{1}{2}.$$

We can express the model for the 2AFC task similarly,

$$p(c)^{\text{2AFC}} = \frac{k}{2N} + \frac{1}{2}.$$

Of course the two above equations are identical. Therefore, choosing the SDT measure d' to compare the two tasks may have incorrectly led to the impression that the 2AFC task was interfering with performance. The above processing models imply that the two tasks can be compared on the basis of proportion correct, but unfortunately raw accuracy data were not reported. Assuming the above described models of performance with categorically distinct stimuli (Rouder et al., 2008; Donkin et al., 2014), at the very least, the analysis of d' *overestimates* the difference between the same-different and 2AFC tasks. Thus our findings may not diverge greatly from the data in Makovski et al. (2010).

Further, our findings are in line with other previous studies that have directly compared the more conventional versions of the change detection task shown in Figure 2.1 and found more or less identical levels of performance (Jiang et al., 2000; Luck & Vogel, 1997; Yang, Tseng, & Wu, 2015). It has been suggested that the contextual support afforded by a whole display compensates for any cost for making multiple at-test decisions (Jiang et al., 2000; Makovski et al., 2010). Although our findings are made all the more surprising by the fact that contextual information will surely have been distorted when items were shuffled between study and test. Indeed Jiang et al. (2000, Experiment 3) compared change detection performance with whole displays in which the relative locations of items were maintained or items were placed in different locations, and thus a different configuration, at test. Destroying the initially presented context led to poorer change detection performance for both colour and shape, however unfortunately these conditions were not compared to a single probe. It would be interesting to make this comparison in further work but nevertheless these findings support the notion that the extra contextual information provided by a whole display is less useful when items move locations between study and test. While contextual information or representation of ensemble properties of the array (see, Brady & Alvarez, 2011) may be distorted by shuffling item locations in our studies it still may benefit performance and offset any cost

associated with making multiple comparisons at test. It will take further, more fine grained studies comparing multiple types of testing procedure to establish whether or not decision load affects change detection performance, however for now there is no strong evidence to suggest that it does.

That the dual probe and whole display procedures do not differ greatly in terms of performance makes sense in the light of other recent findings from research on VWM. The search process engaged in change detection appears to be rather similar for both single probe and whole display tasks. Event related potentials suggest that changes are detected rapidly and attention is oriented to a change at the onset of the whole display test array (Hyun, Woodman, Vogel, Hollingworth, & Luck, 2009). Further, for both probe types, there appears to be an additional, capacity limited, search through VWM (Gilchrist & Cowan, 2014; Hyun et al., 2009). Surprisingly participants seem to engage in this exhaustive search in the standard single probe task, where location of the item can be used to restrict memory search (see also, Z. Chen & Cowan, 2013; Cowan et al., 2013). Thus regardless of whether a single item, two items, or many items must be evaluated at test it seems that observers perform a full search of VWM. It is therefore perhaps unsurprising that we did not find a clear difference between the methods of probing VWM in Experiment 3.

Turning to the main point of these present studies, we find little evidence to support the idea that VWM for feature bindings is specifically disrupted by a whole display probe. The results of Experiments 1 and 2 in terms of the estimated number of items in VWM even suggest the opposite pattern, with binding specific single probe interference. Although there are good reasons, discussed above, to believe that this is the result of problematic averaging over conditions with different logical maximum estimates. The results of Experiment 3, without relying on post hoc modelling, lend no support to the suggestion that bound feature representations are overwritten or disintegrated by a whole display probe.

At first glance our findings appear to conflict with a number of other studies that have suggested, or have been cited in support of the idea, that representations of feature conjunctions are particularly fragile. For example, Logie et al. (2009)

assessed the effect of repeating the same features or conjunctions in memory arrays every three trials (Experiment 1) or on every single trial (Experiment 2) on change detection performance. Surprisingly even when identical conjunctions were presented on every single trial there was no evidence that participants improved at detecting feature swaps across three consecutive test blocks. This is certainly in line with VWM being rather fragile (see also, e.g., Griffin & Nobre, 2003; Makovski, Sussman, & Jiang, 2008) but the data presented do not support the notion that feature bindings are *disproportionately* fragile. Logie et al. (2009) also included trials on which only the colour *or* the shape of their six objects was repeated, given that their task was to detect features swapping location this would be expected to improve performance if learning of the features occurred. This was not the case and their Experiment 2 showed that, if anything, repeating the conjunction of features on every trial led to better overall performance relative to repeating only one of the features. A third experiment, using probed recall, showed that repetition of conjunctions led to learning across trial blocks whereas repetition of features did not. The precise mechanisms underlying these patterns of results are unclear but clearly neither would be expected if bound representations were *particularly* fragile.

One finding that is potentially more difficult to reconcile with the present work is the *visual suffix interference* effect (R. J. Allen et al., 2015; Ueno, Allen, et al., 2011; Ueno, Mate, et al., 2011). Ueno, Allen, et al. (2011) used a similar single probe change detection task to that of Wheeler and Treisman but on some trials presented a to-be-ignored visual suffix (coloured shape) 250 ms after the offset of the memory array. This visual suffix appeared to have a general effect on change detection performance but when the suffix was made up of features that could plausibly come from the task set Ueno, Allen, et al. (2011) observed a small additional deficit for detecting changes to colour-shape binding. Subsequent work has suggested that the suffix overwrites the previous object representations, as participants are likely to erroneously recall features from the to-be-ignored item (R. J. Allen et al., 2015; Ueno, Mate, et al., 2011). This has led to an account of VWM in which features and binding between them are stored at two separate levels (Baddeley, Allen, & Hitch,

2011), similar to the original proposition of Wheeler and Treisman (2002). An attentional filter is said to operate to prevent irrelevant information entering VWM, hence the general effect of the suffix. In addition filtering is said to occur at the object level, thus when this filtering system fails, which is more likely for potentially task-relevant features, the binding between features is lost but the component parts remain at the redundant feature level, hence the binding specific effect.

However, it is important to emphasise that the visual suffix effect occurs with a *single* item. It may be that the general binding cost we observe, relative to feature conditions, is due to probe items in both tasks causing binding specific interference. To our knowledge, no one has assessed whether the number of suffixes modulates the size of the visual interference effect. If this were found to be the case then explaining our pattern of results would be made more difficult. While they did not set out to answer this question, R. J. Allen et al. (2015) recently looked at the suffix interference effect with two to-be-ignored items. Looking at their Figures the binding specific effect appears to be no greater than their previous studies with a single suffix (Ueno, Mate, et al., 2011). Thus it is possible that even our single probe had a specific interfering effect on VWM for bindings, however it is not clear that we would expect this effect to be larger for multiple items. The filter proposed by Baddeley, Allen, and Hitch (2011) could either work in an all-or-none fashion, in which case number of items may not have an effect, or may be put under greater strain when more locations must be filtered, in which case more suffixes would have more of an effect. Either seem possible and distinguishing between these two in future work will help us compare our findings with those of Ueno and colleagues.

So what is the best way to probe VWM? Of course this depends on the question(s) one would like to address. In the present work we have made steps towards better contrasting the different methods of probing VWM however there is still much to learn, especially in the case of the whole display task. Thanks largely to the suggestion of possible whole display interference, the single probe task has been much more widely used and consequently much better characterised (a small selection of such studies includes, R. J. Allen et al., 2006; Bays & Husain, 2008; Donkin et al.,

2013; Cowan et al., 2013, 2014; Rouder et al., 2008), although this is not to say that the whole display task has been completely neglected (e.g., Hyun et al., 2009; Logie et al., 2011, 2009).

An example of a large gap in our understanding of the whole display task is the suggestion that, unlike the single probe task, guessing in this version may be informed by the observer's capacity. As discussed by Rouder et al. (2011), if the observer has access to the number of items in VWM the probability that the participant guesses 'change' given that no-change was detected (i.e. all items in VWM match items in the test array) should be lower for participants with high capacity compared to those with low capacity (see also, R. D. Morey, 2011). Whether or not participants actually engage in this optimal use of information to guide guessing is a fascinating, open question. Also while single probe change detection appears to be in line with the predictions of an all-or-none recognition process, like the one implied by the processing models used here (Donkin et al., 2013, 2014; Rouder et al., 2008), the same has not been established for whole display change detection. It is possible that participants are able to use partial information with a whole display, such as the representation of the ensemble properties of the memory items (Brady & Alvarez, 2011), that will invalidate the use of threshold based models of recognition. Our finding that the whole display task causes little interference is a start and will hopefully encourage more research with this paradigm, as not only does it offer the ability to study guessing behaviour in more detail but it also holds great promise in assessing the comparison of VWM representations to perceptual input (Hyun et al., 2009).

Method for Subsequent Ageing Studies

For the present work—assessing age-differences in forming and temporarily retaining bound features in VWM—we opt for the single probe task. Primarily this is to facilitate comparison with previous research which has hinted at situations where older adults may struggle to retain bindings in VWM (Brown & Brockmole, 2010; Cowan et al., 2006). Also, as discussed above, there are a number of open questions

regarding the whole display task and the way in which information is used in this task, whereas the single probe task is, arguably, better characterised. Further, while Experiment 3 suggests that there is little effect of having to make multiple comparisons at test in younger adults we cannot be certain that this is also the case for healthy older adults. Indeed healthy older adults appear to be less able to filter out irrelevant information in simple VWM tasks (Gazzaley et al., 2008; Jost et al., 2011; Sander, Werkle-Bergner, & Lindenberger, 2011b). Therefore we may expect older adults to experience specific difficulty in processing multiple items at test. Whether or not older adults are specifically affected by decision load at test is an interesting question, but beyond the scope of the present work.

We make use of the processing models derived from the slots view of VWM as a more principled account of performance in the change detection tasks that we use (Cowan et al., 2013; H. Zhang et al., 2010). Further, in Chapter 7 we go beyond the present analysis, in which closed form estimators of k were used, and fit these models to the data directly. This allows us not only to address potential age-differences in k across conditions, but also differences in guessing bias and the propensity to experience lapses of attention during the task (Rouder et al., 2011; R. D. Morey, 2011). Of course, Experiment 3 suggests that the processing models used in Experiments 1 and 2 overestimate feature k relative to bindings. While this is an important issue for studies assessing the *absolute* magnitude of the binding cost it is less of an issue here as we will be assessing the *relative* difference in binding cost between younger and older adults. Further when the processing models are fit directly to data (to extract capacity, guessing, and attentional contributions to performance), as we do in Chapters 4 and 5, the effects of problematic averaging are less pronounced (see, R. D. Morey, 2011).

Chapter 3

Ageing and Feature Binding: The Role of Presentation Time

3.1 Introduction

As outlined in the main introductory section the long-term episodic memory literature has revealed a disproportionate effect of healthy ageing on the ability to form associations *between* distinct items (Old & Naveh-Benjamin, 2008a). This has more recently been traced back to the initial encoding and maintenance of relations in working memory (T. Chen & Naveh-Benjamin, 2012; Hartman & Warren, 2005), however the exact underpinnings of the associative deficit remain unclear (see, Kim & Giovanello, 2011; Naveh-Benjamin et al., 2007; Shing et al., 2008). In contrast to this, the ability to form temporary associations of features *within* objects (e.g., retaining what colour appeared with what shape) appears to be well preserved relative to temporary memory for individual features (Brockmole et al., 2008; Brockmole & Logie, 2013; Isella et al., 2015; Parra, Abrahams, Logie, & Della Sala, 2009; Read et al., 2016).

However, the findings of Brown and Brockmole (2010) suggest that there may be certain conditions under which healthy older adults struggle to form temporary bound representations in VWM. They conducted two experiments examining the role of attentional resources in younger and older adults' ability to bind shape and

colour in VWM. In their first experiment they compared the effect of counting backwards in threes during each change detection trial with a less demanding concurrent articulatory suppression condition. In the second experiment they compared simultaneous and sequential presentation of memory objects, motivated by the suggestion that bindings are more susceptible to interference or overwriting by subsequent items than individual features (R. J. Allen et al., 2006; Logie et al., 2009; Wheeler & Treisman, 2002). In the Brown and Brockmole (2010) study, both the more demanding backwards counting in Experiment 1, and the sequential presentation in Experiment 2 showed evidence of disrupting performance for shape-colour binding to a greater extent than individual features (although see, R. J. Allen et al., 2012). Crucially, in both experiments there was no evidence of a three-way interaction between age, memory condition and attentional manipulation, suggesting that the disruptive effect was similar for younger and older adults. However, a comparison of the two experiments yielded an interesting pattern of results. In Experiment 1 there was no evidence of an age-related binding deficit; that is, there was no significant interaction between age-group and memory condition (shape only, colour only, and shape-colour binding). By contrast, in Experiment 2 there was evidence for an age-related binding deficit in the form of an age by memory condition interaction, with binding showing a larger age effect than individual features alone.

As Brown and Brockmole note, a key difference between the two experiments was the duration for which memory objects were presented. In Experiment 1 the memory array was presented for 900 ms, whereas for Experiment 2 this was increased to 1500 ms for the simultaneous presentation condition in order to equate the presentation time with that used for the sequential condition. One possible explanation for this unexpected finding is that short stimulus exposures may accommodate an *automatic* temporary binding mechanism based largely on early perceptual processing. On the other hand, longer exposures may allow for the deployment of general attentional resources to process and elaborate on the different feature combinations present in an array (R. J. Allen et al., 2006, 2012; Mitchell, Johnson, Raye, Mather, & D'Esposito, 2000). Studies assessing the role of general attentional (or execu-

tive) resources in feature binding have tended to use short stimulus exposures (< 1 second) and have consistently shown that VWM for feature bindings is no more impaired by demanding concurrent tasks than VWM for individual features (e.g., R. J. Allen et al., 2006, 2012; J. S. Johnson et al., 2008; C. C. Morey & Bieler, 2013; Yeh et al., 2005). On the other hand an unpublished study by Elsley and Parmentier (cited in Elsley & Parmentier, 2009) presented memory objects for 2000 ms and found that concurrent maintenance of words disrupted VWM for colour-shape bindings.

While this work is far from conclusive it is suggestive of a greater role for general attentional resources in temporary feature binding in VWM when stimulus exposure is extended. Older adults often exhibit deficits on tasks requiring effortful or controlled processing whereas tasks relying on relatively automatic processes are largely spared (Craik & Bialystok, 2006; Craik & Byrd, 1982). Therefore, if the formation of integrated representations becomes more demanding of attention (more *'active'*, R. J. Allen et al., 2006) with extended presentation time, it is conceivable that older adults are less able to make use of the extra time (e.g., Craik & Rabinowitz, 1985).

Establishing the circumstances under which temporary feature binding in VWM is age-invariant or not is not only theoretically important but also has practical implications given the pronounced binding deficit observed in early Alzheimer's disease (e.g., Della Sala et al., 2012; Parra, Abrahams, Fabi, et al., 2009; Parra, Abrahams, Logie, Mendez, et al., 2010). Identifying boundary conditions where healthy older adults struggle to retain simple feature combinations in VWM can only improve the sensitivity of this task to pathological ageing (Didic et al., 2011; Parra, 2014).

The effect of presentation time on older adults' ability to bind features was not of direct interest to Brown and Brockmole's (2010) experimental manipulations. Consequently, the comparison was made between experiments, that is, between participants, and across two different experimental paradigms. A within participant comparison across directly comparable experimental conditions would make for a stronger test. Therefore, the present study set out to directly assess the effect of presentation time on younger and older adults' ability to bind the shape and

colour of objects in VWM. In their first experiment, Brown et al. (2016) failed to find an effect of presentation time on older adults' binding performance using the same durations as Brown and Brockmole (900 and 1500 ms; Brown et al., 2016). Therefore, we decided to opt for a longer presentation time (2500 ms) in order to increase our chance of finding an age-related effect if one exists.

3.2 Experiment 4 – Presentation Time and Age-Differences in Binding Performance

Method

Participants

Twenty-four younger adults (15 female), aged 18–25 ($M = 21.37$, $SD = 2.10$), were recruited from the student population of the University of Edinburgh and were given either course credit or £5 in return for participation. The older adult group comprised 24 members (16 female) of the University of Edinburgh, Psychology research volunteer panel drawn from the local community, aged 67–78 ($M = 73.17$, $SD = 3.69$), each given £5 in return for participation. Prior to participating in the main experiment all older adults completed the Mini Mental State Examination (MMSE: Folstein, Folstein, & McHugh, 1975) and both age-groups completed the National Adult Reading Test (NART: Nelson, 1982) in order to obtain an estimate of verbal IQ. Normal colour vision was confirmed using a colour blindness test (Dvorine, 1963). All older adults scored 27 or above on the MMSE ($M = 29.46$, $SD = 0.93$). Predicted verbal IQ scores from the NART were clearly higher in the older group ($M = 120.18$, $SD = 5.06$) relative to the younger group ($M = 108.40$, $SD = 5.44$). Years of education on the other hand did not differ greatly between groups (Older: $M = 16.25$, $SD = 3.28$; Younger: $M = 16.15$, $SD = 2.18$).

Stimuli and Apparatus

In line with the experiments of Brown and Brockmole (2010) memory arrays consisted of three coloured shapes presented on a grey background. Each object in the memory array was constructed by combining one of six colours (blue, green, purple, red, turquoise, and yellow) with one of six shapes (arrow, diamond, circle, cross, heart, and triangle), randomly without replacement. Test arrays consisted of a single probe, the nature of which differed according to the memory condition (colour, shape, or binding). When assessing VWM for colour the test object was a 'blob' shape filled in a single colour. For blocks assessing shape memory the test item was a black outline of a shape filled in to match the background. Finally for binding trials the test object was a coloured shape (see Figure 3.1). Stimuli were presented on a 22" LCD monitor. Each object measured approximately 2 cm × 2 cm which at a viewing distance of approximately 57 cm corresponds to 2° × 2° visual angle. Objects in the memory array were presented in a row separated centre-to-centre by approximately 5 cm and were centred 3 cm above a central fixation cross. Test items appeared in analogous positions 3 cm below the central fixation. The location occupied by the test item was chosen at random. The experimental sequence was implemented in E-Prime (Schneider, Eschman, & Zuccolotto, 2002).

Procedure

The general trial procedure used in the change detection task is shown in Figure 3.1. Participants initiated each trial, when ready to do so, by pressing the spacebar on the keyboard. They were then presented with a randomly generated number, between 20 and 99, for 2 seconds which they were required to repeat aloud at a steady pace throughout the trial, until the response was made. The experimenter recorded the number of articulations made on each trial and ensured a stable rate of articulation. Following the number a blank central fixation screen was presented for 1 second and the fixation cross remained visible throughout the trial. The memory array was then presented for 900 or 2500 ms depending on the current block. Following a 1 second retention interval the probe item was presented and remained visible until

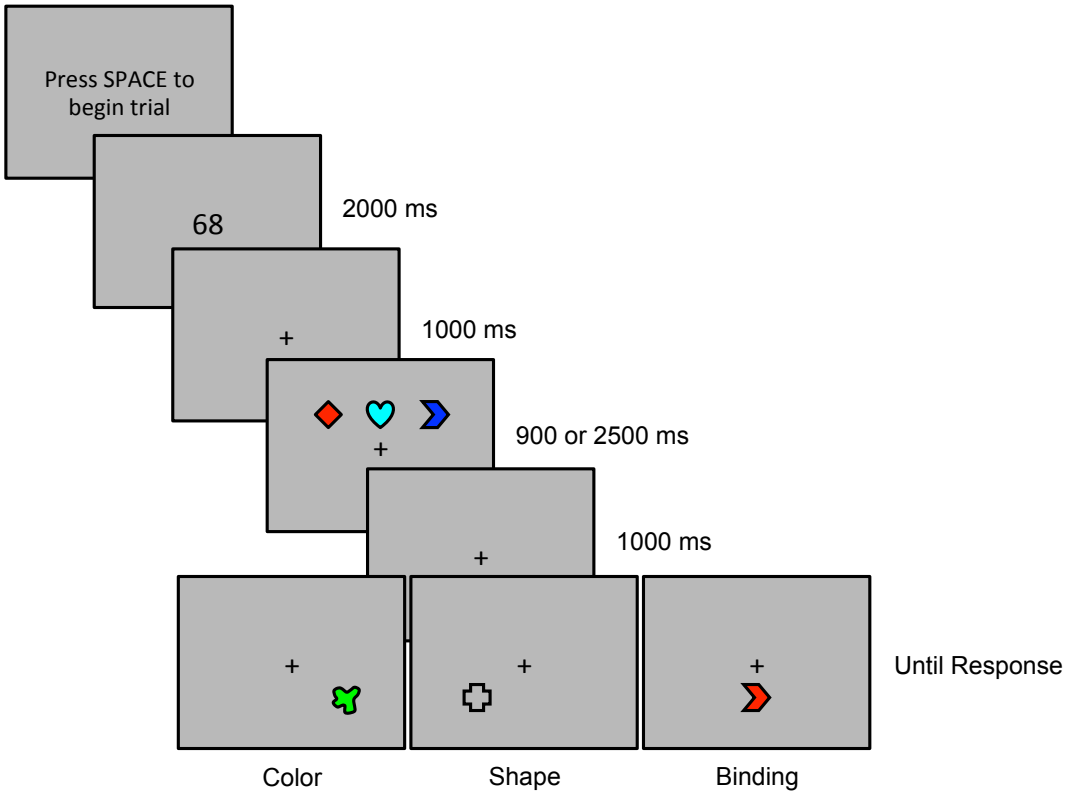

Figure 3.1: Trial sequence during change detection task for colour, shape, and binding in Experiment 4. *Note:* No-change trials not depicted and items are not drawn to scale.

the response was made. Participants were required to indicate whether the test item had appeared in the previous memory array or if a change had occurred by pressing either the *'z'* key (labelled *'YES'*) or the *'m'* key (labelled *'NO'*), respectively. For change trials in the individual feature conditions (colour or shape only) the test object was randomly selected from the three remaining colours or shapes not present in the memory array. For binding change trials the test item was created by recombining a shape and colour from the initial memory array that had not appeared together (see Figure 3.1 for an example). This 'feature swap' method ensures that participants are required to remember the binding of shape and colour to detect a change rather than the features individually (cf. Chalfonte & Johnson, 1996).

The experiment was divided into 6 blocks combining the 2 presentation times and 3 memory conditions. Participants completed all memory conditions at a given presentation time before moving on to the next. Half of the participants in each

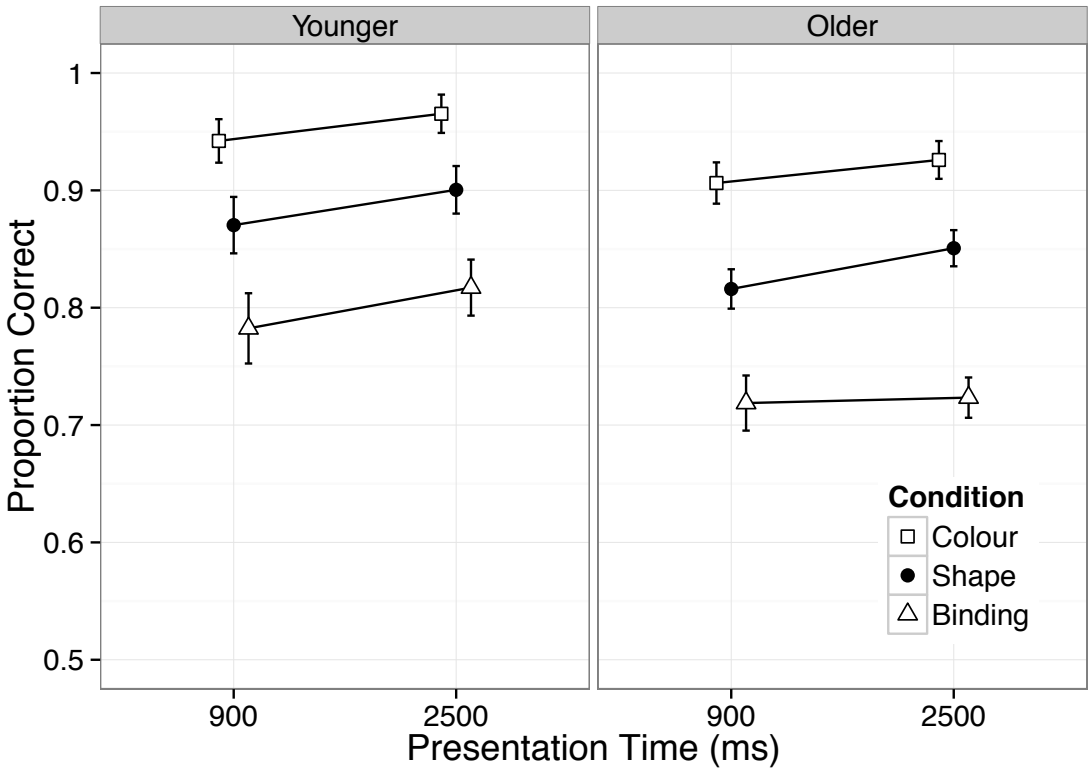

Figure 3.2: Proportion correct for Experiment 4 comparing the effect increasing presentation time on VWM for feature bindings in younger and older adults. Error bars are ± standard error.

age-group completed the 900 ms condition first and the other half completed the 2500 ms condition first. The order of memory conditions was counterbalanced with the constraint that each participant completed the memory tasks in the same order at each presentation time. Each block began with 6 practice trials followed by 36 experimental trials with breaks provided between blocks. In each block 50% of trials were change trials and 50% were no-change trials.

Results

Proportion correct

The mean proportion correct responses given by each age group in each condition is presented in Figure 3.2. As noted in Chapters 1 and 2 analysis of aggregated proportions with ANOVA can lead to erroneous conclusions, especially regarding interactions which are of primary interest here (Dixon, 2008; Jaeger, 2008). Thus

Table 3.1: Posterior quantities from logit model for Experiment 4

			95% HDI		
Parameter	Mean	Median	lower	upper	ESS
β_0	2.035	2.034	1.848	2.214	1704.010
$\boldsymbol{\beta}_1$: (1) Shape	-0.077	-0.077	-0.165	0.010	23089.035
$\boldsymbol{\beta}_2$: (2) Binding	-0.787	-0.787	-0.868	-0.706	17792.203
$\boldsymbol{\beta}_3$: (3) 2500 ms	0.140	0.140	0.074	0.205	18591.099
$\boldsymbol{\beta}_4$: (4) Older Group	-0.319	-0.319	-0.495	-0.136	1736.110
$\boldsymbol{\beta}_5$: 1×3	0.006	0.006	-0.081	0.094	22114.129
$\boldsymbol{\beta}_6$: 2×3	-0.074	-0.074	-0.155	0.006	18445.085
$\boldsymbol{\beta}_7$: 1×4	0.039	0.039	-0.047	0.128	22780.903
$\boldsymbol{\beta}_8$: 2×4	0.063	0.062	-0.017	0.147	18364.236
$\boldsymbol{\beta}_9$: 3×4	-0.050	-0.049	-0.115	0.016	18496.069
$\boldsymbol{\beta}_{10}$: $1 \times 3 \times 4$	0.033	0.033	-0.054	0.121	22409.320
$\boldsymbol{\beta}_{11}$: $2 \times 3 \times 4$	-0.005	-0.005	-0.086	0.076	19091.858
σ_s	0.587	0.581	0.454	0.727	13845.274

Note: The effects coded variables were as follows: (1) Shape = 1, Binding = 0, Colour = -1, (2) Shape = 0, Binding = 1, Colour = -1, (3) 900 ms = -1, 2500 ms = 1, (4) Younger = -1, Older = 1. Interaction contrasts were products of these effects coded variables.

we estimated the logit model described in Chapter 2 with effects coded variables reflecting age, condition, and presentation time. The resulting parameter estimates (see Table 3.1) suggest that correct responses were more likely following longer exposure, 0.281 [0.148, 0.411]. Older adults were less likely to respond correctly overall, -0.638 [-0.989, -0.273], and performance in the individual feature conditions was much better than that in the binding condition, 1.181 [1.059, 1.302].

Turning to interactions; there was some indication that the disparity between the feature and binding conditions was less pronounced at the longer presentation time, 0.223 [-0.017, 0.465], although this specific contrast is not credibly different from zero. There was also some suggestion that older adults benefited less from the increase in exposure duration, -0.198 [-0.459, 0.063], although, again, as the HDI overlaps zero it is difficult to make a strong statement regarding this specific contrast.

For the crucial interaction between age and condition the contrast between features and bindings[1] tends towards a smaller binding cost in the younger group, -0.188

[1]It is common to compare binding change detection to performance in blocks assessing memory

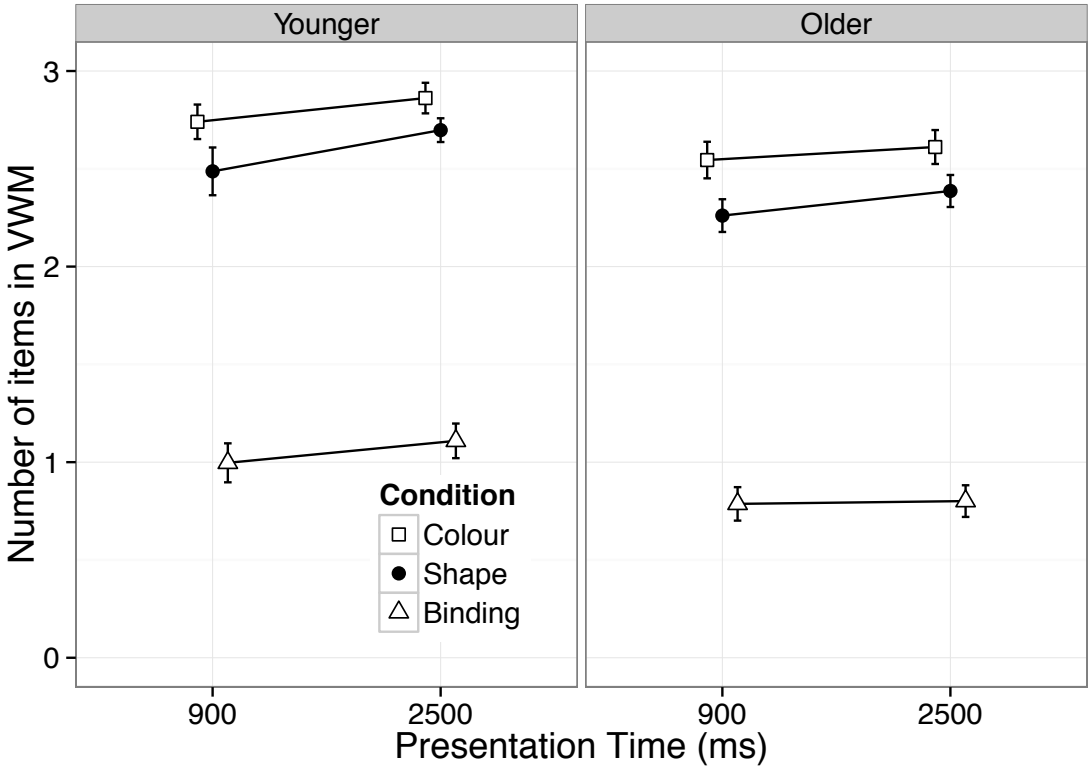

Figure 3.3: Estimated number of items in VWM for Experiment 4. Error bars are ± standard error.

[-0.440, 0.051]. However, this is clearly a very small effect that cannot be credibly distinguished from zero. Finally there was no suggestion that this was modulated by presentation time, 0.029 [-0.459, 0.514], in contrast to the findings of Brown and Brockmole (2010). In order to look at these trends in more detail, and to try and assess the evidence *against* these crucial interactions, we also looked at model-based estimates of the number of items in VWM.

Number of items

The number of items in VWM was estimated from hit and false alarm rates using the appropriate formulae for this single probe task described in Chapter 2. Figure 3.3 presents these estimates across groups and conditions. Table 3.2 shows that the winning model from our default Bayes factor analysis included main effects of

for shape only (e.g. Brockmole et al., 2008). However, we think it is more informative to compare binding performance to the average of individual feature performance, thus using all the information gathered. Adopting this approach does not change crucial conclusions unless otherwise noted.

Table 3.2: Log Bayes factors for Experiment 4

	Model	$\log(B_{M,0})$	% error
1	$k \sim$ C + PT + AG + ID	241.42	0.71
2	$k \sim$ C + PT + AG + PT:AG + ID	240.13	0.71
3	$k \sim$ C + AG + ID	239.97	0.43
4	$k \sim$ C + PT + ID	239.50	0.97
5	$k \sim$ C + PT + C:PT + AG + ID	239.24	0.97
6	$k \sim$ C + PT + AG + C:AG + ID	238.82	0.53
7	$k \sim$ C + ID	238.05	0.24
8	$k \sim$ C + PT + C:PT + AG + PT:AG + ID	237.97	1.21
9	$k \sim$ C + PT + AG + C:AG + PT:AG + ID	237.55	0.96
10	$k \sim$ C + AG + C:AG + ID	237.36	1.12
11	$k \sim$ C + PT + C:PT + ID	237.31	0.98
12	$k \sim$ C + PT + C:PT + AG + C:AG + ID	236.66	1.00
13	$k \sim$ C + PT + C:PT + AG + C:AG + PT:AG + ID	235.35	2.64
14	$k \sim$ C + PT + C:PT + AG + C:AG + PT:AG + C:PT:AG + ID	233.25	1.11
15	$k \sim$ AG + ID	-0.15	0.47
16	$k \sim$ PT + ID	-1.53	0.74
17	$k \sim$ PT + AG + ID	-1.68	0.34
18	$k \sim$ PT + AG + PT:AG + ID	-3.33	0.98

Note: All Bayes factors are relative to a null model with a random participant effect only (model 0: $k \sim$ ID). PT = Probe Type, C = Condition, AG = Age Group, ID = participant ID, and ':' denotes an interaction effect

condition, presentation time, and age group, but no interactions. Comparing models we find overwhelming evidence for the effect of condition ($B_{1,17} = 3.759542 \times 10^{105}$) and substantial evidence in favour of the effects of presentation time ($B_{1,3} = 4.26$) and age-group ($B_{1,4} = 6.81$).

For the interactions of primary interest, comparing models 1 and 6 we find good evidence for the *omission* of the age by condition interaction from the winning model ($B_{1,6} = 13.46$). Finally for the three way interaction the comparison of models 13 and 14 shows that a model omitting this interaction is more likely given the data ($B_{13,14} = 8.16$). Thus the results of this Bayes factor analysis strengthen our above analysis of raw accuracy in providing no evidence for (even evidence against) a specific feature binding deficit in healthy older adults.

Discussion

The findings of Brown and Brockmole (2010) presented the possibility that greater time to encode memory items may lead to the appearance of a robust age-related binding deficit. This experiment tested this directly and found no evidence that increasing presentation time led to the emergence of this effect. In fact we were able

to provide substantial-to-strong evidence against the suggestion that age disproportionately affects VWM for feature bindings when assessing model-based estimates of k. We return to this issue in more detail in the General Discussion.

Scrutinising Figure 3.2 we see that performance was particularly poor for *both* groups in the binding condition. Why was there such a disparity between feature and binding performance here relative to other experiments using the single probe paradigm (e.g., R. J. Allen et al., 2006; Brown & Brockmole, 2010)?

One potential explanation again lies in the distinction between automatic and effortful forms of temporary binding (R. J. Allen et al., 2006, 2012; Mitchell, Johnson, Raye, Mather, & D'Esposito, 2000). If greater time to study memory items leads to more elaborative forms of encoding these complex processes may be more volatile and likely to fail (which may help explain the present pattern) or they may lead to strengthened object representations, closing the gap between the salient individual features and feature bindings (in opposition to our large binding cost). Even our shorter presentation time was quite lengthy relative to other similar studies of feature binding (e.g., R. J. Allen et al., 2006; Cowan et al., 2013; Delvenne et al., 2010) and given that, to our knowledge, no one has examined the effect of exposure duration on the binding cost we conducted a second experiment with younger adults in which items were presented for only 200 ms.

3.3 Experiment 5 – Shorter Presentation and the Binding Cost

Method

Participants

Twenty-four younger adults, aged 18-29 ($M = 22.11$, $SD = 2.48$), who had not taken part in Experiment 4 were recruited from the student population of the University of Edinburgh and were given either course credit or £5 in return for participation. As in Experiment 4, prior to the main experiment participants completed the NART

Table 3.3: Posterior quantities from logit model for Experiment 5 along with the same analysis for the younger adults in Experiment 4

	Experiment 5				Experiment 4 - younger only			
Parameter	Mean	Median	lower	upper	Mean	Median	lower	upper
β_0	2.059	2.057	1.801	2.321	2.362	2.361	2.043	2.685
$\boldsymbol{\beta}_1$: (1) Shape	-0.227	-0.227	-0.347	-0.103	-0.116	-0.116	-0.254	0.020
$\boldsymbol{\beta}_2$: (2) Binding	-0.772	-0.771	-0.888	-0.659	-0.855	-0.854	-0.983	-0.729
$\boldsymbol{\beta}_3$: (3) Long	0.083	0.082	-0.012	0.175	0.190	0.190	0.087	0.293
$\boldsymbol{\beta}_4$: 1×3	0.039	0.040	-0.084	0.160	-0.026	-0.026	-0.165	0.108
$\boldsymbol{\beta}_5$: 2×3	-0.055	-0.055	-0.170	0.060	-0.069	-0.069	-0.196	0.054
σ_s	0.577	0.565	0.385	0.789	0.747	0.733	0.523	1.004

Note: The effects coded variables were as follows: (1) Shape = 1, Binding = 0, Colour = -1, (2) Shape = 0, Binding = 1, Colour = -1, (3) Short = -1, Long = 1. For Experiment 5 Short = 200 ms and Long = 900 ms. For Experiment 4 Short = 900 ms and Long = 2500 ms. For all deflection ($\boldsymbol{\beta}$) parameters ESS > 10000.

and colour vision was confirmed using the colour blindness test (Dvorine, 1963). This group had a mean (SD) predicted verbal IQ score of 109.73 (4.29) and 16.17 (1.90) years of education, neither of which differed greatly from the younger group in Experiment 4.

Stimuli and Procedure

The stimuli and procedure were identical to Experiment 4 with the only exception being that in the shorter presentation time condition the memory array appeared for 200 ms and in the longer condition for 900 ms.

Results

Proportion correct

Table 3.3 presents the outcome of the logit model fit to the raw accuracy data from Experiment 5 as well as the same analysis for the younger adults from Experiment 4 for comparison. Looking at specific contrasts we find that correct answers were much more likely for features relative to bindings, 1.157 [0.988, 1.332]. This effect is of a similar magnitude to that observed in the younger sample of Experiment 4, 1.282 [1.094, 1.475]. There was no clear effect of presentation time in this experiment, 0.165 [-0.024, 0.350], in contrast to Experiment 4, 0.380 [0.174, 0.585]. Finally there

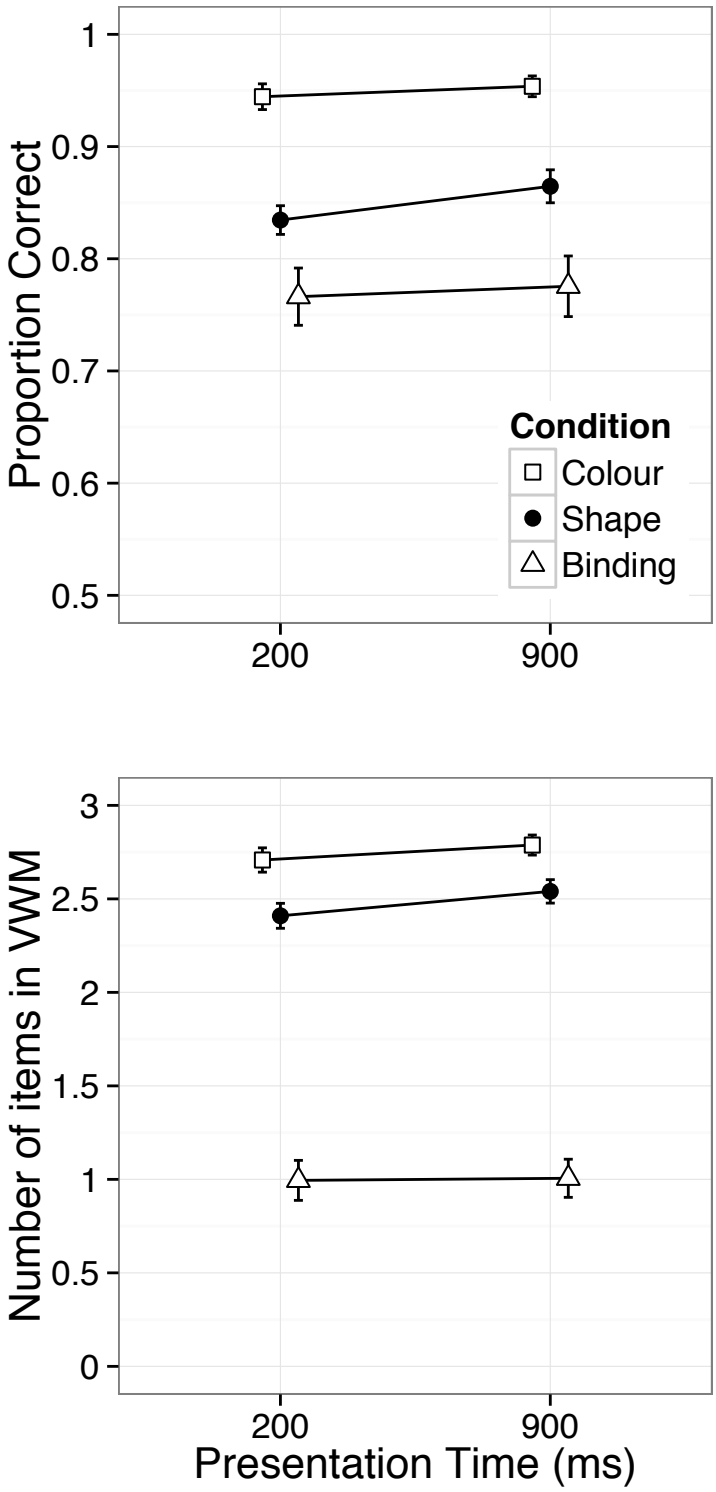

Figure 3.4: Data from Experiment 5 across condition and presentation time. Upper panel presents proportion correct and lower presents the estimated number of items in VWM. Error bars are ± standard error.

was no suggestion in the data of a substantial interaction between presentation time and the difference between features and binding, 0.166 [-0.181, 0.509]. This interaction contrast is very similar to that observed for younger adults in Experiment 4, 0.207 [-0.163, 0.589].

Number of items

For the estimated number of items in VWM the default Bayes factor analysis revealed a winning model containing only a main effect of condition ($\log(B_{1,0}) = 132.96$). Comparing to the second model which contained condition and presentation time ($\log(B_{2,0}) = 132.31$) we find that the omission of presentation time is favoured by approximately 2-to-1. Model 3 contained the interaction between condition and presentation time ($\log(B_{3,0}) = 130.57$). Contrasting model 2 and model 3 we find substantial evidence against the two-way interaction ($B_{2,3} = 5.7$).

We conducted a similar analysis with the data from the younger group in Experiment 4. The winning model included both main effects without the interaction ($\log(B_{1,0}) = 127.05$). Model 3 omitted presentation time ($\log(B_{3,0}) = 124.9$) and compared to model 1 we find that the presence of this effect is favoured by around 9-to-1. Thus while the evidence against an effect of presentation time from Experiment 5 is not convincing it is clearly the case that the effect of increasing from 900 to 2500 ms was greater than increasing from 200 to 900 ms, in line with our above analysis. The second best model included the interaction ($\log(B_{2,0}) = 125.21$) and contrasted with the winning model there was weak evidence for its omission ($B_{1,2} = 1.92$).

Discussion

Comparing the left panel of Figure 3.2 to the top panel of Figure 3.4 there is very little difference, beyond the overall level of performance in the shortest presentation condition. The analysis backs this up in showing that exposure duration did not modulate the disparity between feature and binding performance. Thus it is still

unclear as to why binding performance, relative to features, was so poor in these experiments.

Since the original study of Wheeler and Treisman (2002) in which they showed that, with a single probe, performance in shape only and binding conditions does not significantly differ it has been argued that the difference between these conditions is an index of the efficacy of feature binding (see, e.g., R. J. Allen et al., 2006). However, a detailed look at the literature reveals that performance differences between feature and binding conditions have not been entirely consistent. Since Wheeler and Treisman (2002) several studies have recreated the finding that performance in the binding condition is roughly equal to (or even better than) the most difficult individual feature condition (R. J. Allen et al., 2006, Experiments 1 and 2; Delvenne & Bruyer, 2004; Delvenne et al., 2010). However, many others using very similar stimuli and the single probe task find small but significant binding costs (R. J. Allen et al., 2006, Experiments 3–5; R. J. Allen et al., 2012; Fougnie & Marois, 2009). There is a tendency for studies that find the binding cost to have larger sample sizes (24 versus < 16) but beyond that there is no clear reason as to why the cost was so large in the present experiments. What is clear, however, is that this is not greatly modulated by presentation time.

3.4 General Discussion

In contrast to the associative deficit seen in healthy ageing (Old & Naveh-Benjamin, 2008a), the ability to form temporary bindings between simple features appears relatively robust across the lifespan (Brockmole & Logie, 2013). However, Brown and Brockmole (2010) reported two experiments which, taken together, suggested a role for increased presentation time in the emergence of an age-related feature binding deficit. This could plausibly be linked to a greater role for attentional resources to engage in a more active form of feature binding at longer stimulus durations (e.g., R. J. Allen et al., 2006), something older adults may struggle with (Craik & Bialystok, 2006). We therefore directly assessed the effect of presentation time on younger and older adults' ability to bind the colour and shape of objects in

VWM.

Our analyses of both proportion correct and the estimated number of items in VWM provide no reason to suspect that healthy ageing disproportionately affects temporary feature binding or that increasing exposure duration changes this (see S. Rhodes, Parra, & Logie, 2016, for analysis of additional measures). On the contrary, models omitting these interaction effects were more likely given the data (and the default priors) than models including them.

It may be that small demographic differences between the sample recruited for the present study and that of Brown and Brockmole, especially their Experiment 2, account for the absence of the age-group interaction. Our sample had a slightly higher mean years of education compared to Brown and Brockmole's sample in their Experiment 2 (16.25 versus 13.81) and also obtained a higher mean estimate of verbal IQ from the NART (120.18 versus 115.63). Despite this, we suspect that these slight differences in years of education and verbal IQ cannot account for the absence of the crucial interaction, especially given that these characteristics were well matched between the two experiments of Brown and Brockmole (2010), one of which did find a binding deficit. Moreover, recent studies of VWM binding in populations with different demographic features and health status confirmed that age and education did not yield significant differences between control participants nor did they impact on performance in affected individuals (Parra, Della Sala, et al., 2011). Our conclusion, that increasing presentation time does not lead to an age-related colour-shape binding deficit, is also strengthened by another recent study assessing the effect of presentation time on older adults' binding performance using identical durations to Brown and Brockmole (2010) which also failed to find an age-related binding deficit (Brown et al., 2016, Experiment 1). Further, the present experiment increased the disparity between the shorter and longer presentation times and therefore was, arguably, more likely to find an effect of presentation time.

Explaining why Brown and Brockmole (2010) did find evidence of an age-related binding deficit in their second Experiment is a difficult task. While we suspect differences in verbal IQ and years of education are insufficient to explain this there

remain other sample characteristics that may contribute to the appearance of a binding deficit. As detailed in the Introduction a specific colour-shape binding deficit appears to be a marker of early Alzheimer's disease (Parra, Abrahams, Fabi, et al., 2009), and has even been observed in a familial variant of the disease approximately 10 years before other symptoms of the disease become apparent (Parra, Abrahams, Logie, Mendez, et al., 2010). Whether a random sample of healthy older people show a binding impairment in the group aggregate score would then depend on how many might be at risk for developing dementia, even if they are otherwise asymptomatic at the time of testing. This is an hypothesis that we plan to address in our future research assessing feature binding in mild cognitive impairment. However, it appears clear that in most groups of healthy older adults any binding impairment in temporary memory is either not present or too small to be statistically reliable (Brockmole et al., 2008; Brockmole & Logie, 2013; Isella et al., 2015; Parra, Abrahams, Logie, & Della Sala, 2009; Read et al., 2016), and may be one of the cognitive abilities that is relatively well preserved across the healthy adult lifespan (for a review see Logie et al., 2015).

Another possibility is that the choice of outcome measure plays an important role in the emergence of an age-related binding deficit. L. A. Brown kindly shared the data from Brown and Brockmole (2010) and we performed additional analyses. Looking at proportion correct, rather than A', in a standard analysis of variance we find that the age by condition interaction is significant ($p < 0.01$) but the estimate of partial eta squared ($\eta_P^2 = 0.127$) is around 60% the size of the estimate for A' ($\eta_P^2 = 0.210$). Thus while analysis of proportion correct instead of A' would not have changed the overall conclusions in terms of the 'significance' of the crucial interaction effect its strength is certainly reduced using this measure. R. J. Allen et al. (2012) made similar observations when assessing the effect of concurrent tasks on feature binding in VWM as A' appeared to be more likely to yield significant interaction effects relative to other outcome measures. This illustrates the crucial role that the choice of outcome measure plays in the outcome of cognitive ageing research. We assess the effects of choosing between different outcome measures (d',

A', $p(c)$) on the results of studies such as this one with a series of simulations in Chapter 8.

Of course analysing proportional data with statistical models that assume normally distributed data poses a number of problems, as discussed in Chapter 1 (see also, Dixon, 2008; Jaeger, 2008). Therefore we also fitted a logit model to Brown and Brockmole's Experiment 2 data, using the `lme4` package (Bates, Maechler, Bolker, & Walker, 2014), which revealed no clear interaction between age-group and the difference between feature and binding performance ($\beta = -0.139$, $\text{SE}_\beta = 0.109$, $p \approx 0.2$). Therefore, in hindsight, the initial evidence for an age-related binding deficit provided in Brown and Brockmole (2010) appears to have been an artifact.

In summary we assessed the effect of increasing study time for a change detection task on younger and older adults' ability to form bound temporary representations in VWM. The amount of time given to participants did not differentially affect their ability to detect binding changes relative to changes of individual features. This is in line with a growing body of evidence showing that the ability to the bind surface features of objects in VWM is largely unaffected by age.

Chapter 4

Ageing and Feature Binding: Mixed versus Blocked Trials

4.1 Introduction

As previously discussed and demonstrated in Chapter 3, studies assessing the effects of healthy ageing on the ability to associate surface features in VWM often find no evidence of an age-related binding deficit (Brockmole et al., 2008; Brockmole & Logie, 2013; Bopp & Verhaeghen, 2009; Isella et al., 2015; Parra, Abrahams, Logie, & Della Sala, 2009; Read et al., 2016; S. Rhodes et al., 2016). Another feature that these studies share is that feature and binding changes are presented in *separate* trial-blocks. In the visuo-spatial binding literature Cowan et al. (2006) found that, like younger adults, older adults were just as sensitive to changes of colour-location binding (in terms of d') as they were to colour changes provided that these trials were presented in separate blocks. However, when these different changes were *mixed* together older adults were less sensitive to binding changes. Cowan et al. (2006) suggested that older adults were more likely to perform the change detection discrimination on the basis of probe familiarity which, in the mixed condition, would be sufficient to detect salient feature changes but not the less salient swaps of colour and location. Relying only on familiarity in the blocked condition, on the other hand, would not support the detection of any changes, thus the older adults

in this condition may have adopted encoding or retrieval strategies that improved their performance in the binding condition. For example, as we outline in more detail below, binding trials in Cowan et al's experiments involved the introduction of duplicated colours in the test array so it is possible that participants noticed this when these trials were presented in a separate block and used this to their advantage by focusing on the presence of duplicates in the memory array.

Another way of conceptualising this, as Cowan et al. (2006) did in a signal detection theory analysis, is that older adults use a more conservative response criterion relative to younger adults. Assuming that the signal arising from feature changes is larger than that for binding changes—as implied by commonly observed differences in performance between these conditions (see Chapter 3)—a more conservative criterion would lead to an increased frequency of misses for binding changes. Differences in sensitivity for binding changes would increase this problem further. Mixing trial types, as opposed to presenting them in separate trial blocks, may also reveal strategy differences at encoding given that in the blocked condition participants know what kind of test to prepare for.

However, the findings of T. Chen and Naveh-Benjamin (2012) cast doubt on the role of mixing trials in the emergence, or exacerbation, of age-related binding deficits. They used a continuous recognition paradigm in which participants studied a stream of face-scene pairs with interspersed tests of memory for items or associations following varying delays. The commonly observed associative deficit (see Chapter 1) was no larger when item and associative trials were mixed together relative to when they were encountered in separate trial blocks. Thus the role of mixing or blocking trials in producing or magnifying age-related binding deficits is unclear and, given that, to our knowledge, only two studies have directly addressed this question, it would clearly benefit from further investigation. Here we assessed the effect of mixing changes to colour-shape conjunctions together with colour changes and shape changes on older adults' performance on a change detection task. To preview the results, we again find no evidence that healthy ageing disproportionately affects the short-term retention of bindings between surface features. Further

we observe no-difference between the two types of trial-block either in terms of raw accuracy or in terms of an analysis of discriminability and response bias.

4.2 Experiment 6 – Mixed versus Blocked Trials

Method

Participants

Forty-eight younger adults from the student population of the University of Edinburgh took part in return for £5 for the 45 minute session. Forty-nine healthy older adults from the University of Edinburgh Psychology volunteer panel also took part and were offered £5 in return for participation. These groups were split between two conditions; one in which colour, shape and binding changes were mixed together and another in which they were presented in separate trial blocks. Table 4.1 provides participant's demographic information; the two age-groups were roughly equated for years of education and older adults consistently outperformed on the NART. All older adults scored 27 or above on the MMSE.

Table 4.1: Participant characteristics across the mixed and blocked conditions of the present experiment

	Blocked		Mixed	
	Younger	Older	Younger	Older
N	24	24	24	25
N_{Female}	17	17	14	18
Mean Age (SD)	20.71 (2.90)	70.96 (5.61)	21.12 (2.66)	70.28 (4.43)
Years of Education	16.10 (2.61)	17.06 (2.90)	16.52 (2.16)	15.96 (2.43)
NART Verbal IQ	111.92 (5.56)	121.23 (4.26)	111.30 (7.14)	118.69 (5.76)
MMSE	-	29.33 (0.76)	-	29.60 (0.82)

Stimuli and Apparatus

Our first ageing study (reported in Chapter 3) used nameable shapes with articulatory suppression to discourage verbal rehearsal of the items (cf. Brown & Brockmole,

2010). However, anecdotal evidence gained from informal discussion with participants suggested that, in spite of suppression, both younger and older adults were able to verbally rehearse the memoranda. Thus for the present experiment we abandoned articulatory suppression and drew our stimuli from sets of 8 difficult to name colours and abstract polygons taken from (Brockmole et al., 2008; Parra, Abrahams, Logie, & Della Sala, 2009). Items in the memory array were constructed by selecting colours and shapes from these sets without replacement. Stimuli were presented on a grey background in 8 locations surrounding the centre of the screen in an invisible circle (radius = 2.6°). Objects measured approximately 1° visual angle and were separated centre-to-centre by at least 2°. The experiment was programmed using PsychoPy (Peirce, 2007, 2009) and presented over a 18" E96f+SB ViewSonic monitor with a resolution of 1024×768 and refresh rate of 100 Hz.

Design and Procedure

Prior to the main change detection task both groups completed the NART (Nelson, 1982) to obtain an estimate of verbal-IQ (see Table 4.1) and a test of colour vision (Dvorine, 1963). The older group also completed the MMSE (Folstein et al., 1975) prior to completing the main part of the experiment (see Table 4.1).

The general trial sequence of the main change detection task is presented in Figure 4.1 along with examples of the kinds of trials presented to participants. Participants initiated each trial by pressing the spacebar and following a 1000 ms fixation cross the memory array appeared for 900 ms. This was followed by a 1000 ms blank retention interval and then a single central probe item which remained present until a response was made. Finally, in line with the procedure of Cowan et al. (2006), participants were presented with feedback for 1000 ms in the form of a fixation cross that was coloured green for a correct response and red otherwise.

Half of the trials presented to participants involved no-change as the probe was selected at random from one of the 3 or 6 objects presented. The remaining half of trials were split between colour change, shape change, and binding change types (either blocked or mixed). A colour change involved filling a previously seen shape

Figure 4.1: Trial sequence and illustrations of different trial types in Experiment 6. The different kinds of change trial were either blocked with no-change trials or were mixed within the same trial-block.

with a colour from outside the original memory set and a shape change involved presenting a new shape in a previously seen colour. A binding change involved presenting a combination of a colour and shape from separate memory objects as the probe item (see Figure 4.1). As described above participants in the blocked condition saw these changes in separate blocks with their own no-change trials, whereas participants in the mixed condition saw the three kinds of change trial interspersed unpredictably with no-change trials.

The main experiment was split into 3 blocks with 32 change and 32 no-change trials distributed evenly across the different set sizes. For the blocked condition all change trials were of a single type and for the mixed condition a change was equally likely to occur for colour, shape, and binding. Participants in the blocked condition were given 6 practice trials looking for a particular kind of change before the corresponding block whereas participants in the mixed condition were given 18 practice trials before the first block with all three kinds present. In the blocked condition the order of the three memory conditions (colour, shape, binding) was fully counterbalanced.

Results

The results of this experiment are presented in two main sections; in the first we present raw accuracy on the change detection task in the blocked and mixed conditions. As raw accuracy reflects both the sensitivity and bias of an observer, in the second section we attempt to separate out these two contributions using measures derived from the two-high threshold model of recognition (Snodgrass & Corwin, 1988).

Raw Accuracy

Blocked Trials The blocked condition involved manipulation of three factors; 1) memory condition (colour, shape, binding), 2) set size, and 3) whether or not a change occurred. Figure 4.2 presents accuracy across these experimental factors and age-groups.

The hierarchical logistic regression model described in Chapter 2 was estimated using raw accuracy and effects coded variables representing main effects and interactions in the design. Using the resulting parameter estimates (see Table 4.2) we constructed specific contrasts to test hypotheses about the data. Correct responses were much less likely when participants had to remember 6 items relative to 3, -0.774 [-0.875, -0.677], and overall older adults were less accurate than younger adults, -0.537 [-0.767, -0.309]. There was no clear overall difference between trials on which the probe changed versus trials where no-change occurred, although there was a slight tendency towards better change detection, 0.076 [-0.023, 0.174]. Overall participants were more likely to respond correctly in feature conditions relative to the binding condition, 0.292 [0.191, 0.393].

The discrepancy between overall feature and binding performance was slightly *smaller* in the older group relative to the younger group, -0.112 [-0.318, 0.088]. The 95% HDI for this contrast clearly overlaps with zero but the direction of this comparison is certainly not in line with a specific age-related binding deficit. This crucial two-way interaction did not appear to be modulated by either trial type (change versus no-change), 0.025 [-0.381, 0.424], or set size (6 versus 3), -0.277 [-

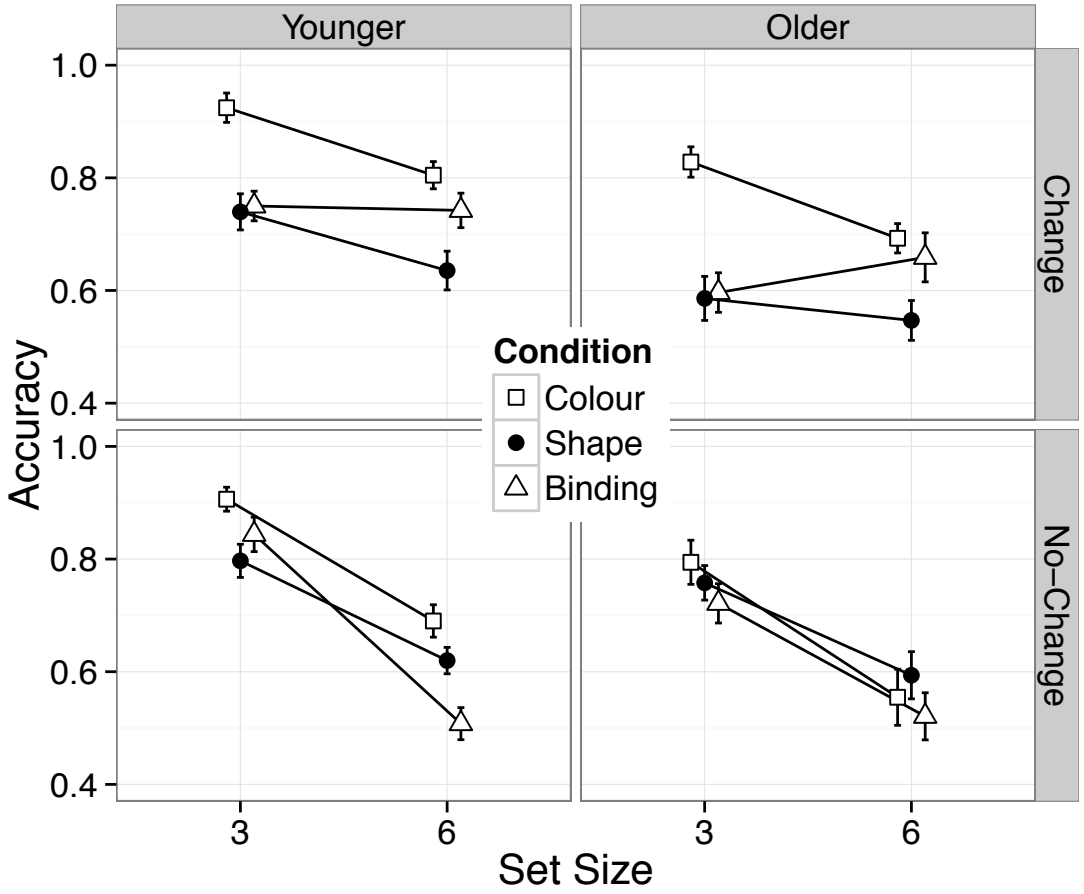

Figure 4.2: Accuracy for blocked trials across age-groups and experimental conditions. Error bars are ± standard error.

0.678, 0.126]. Finally, there was some evidence that the effect of age on accuracy was smaller when no-change was present relative to trials where a change had occurred, -0.204 [-0.401, -0.008].

Thus the results from the blocked condition provide no reason to believe that older adults specifically struggle to detect changes to combinations of colour and shape. We, of course, expected this from the findings of Cowan et al. (2006) and previous assessments of surface feature binding. However, what is less clear is older adults' sensitivity to binding changes when different trial types are mixed together.

Mixed Trials Figure 4.3 displays the pattern of performance when feature and conjunction change trials were mixed together within the same block of trials. Table 4.3 presents the results of our logit model with the experimental factors of trial

Table 4.2: Posterior quantities from logit model for the Blocked condition

			95% HDI		
Parameter	Mean	Median	lower	upper	ESS
β_0	0.960	0.959	0.840	1.075	2780.328
$\boldsymbol{\beta}_1$: (1) Shape	-0.252	-0.252	-0.319	-0.186	19687.658
$\boldsymbol{\beta}_2$: (2) Binding	-0.194	-0.194	-0.262	-0.127	19149.694
$\boldsymbol{\beta}_3$: (3) SS6	-0.387	-0.387	-0.437	-0.338	25007.410
$\boldsymbol{\beta}_4$: (4) Older Group	-0.269	-0.269	-0.383	-0.154	2724.668
$\boldsymbol{\beta}_5$: (5) Change	0.038	0.038	-0.011	0.087	24298.394
$\boldsymbol{\beta}_6$: 1×3	0.092	0.092	0.027	0.159	20216.965
$\boldsymbol{\beta}_7$: 2×3	0.091	0.091	0.023	0.159	19863.237
$\boldsymbol{\beta}_8$: 1×4	0.086	0.086	0.021	0.154	19359.511
$\boldsymbol{\beta}_9$: 2×4	0.037	0.037	-0.029	0.106	19492.504
$\boldsymbol{\beta}_{10}$: 1×5	-0.198	-0.197	-0.265	-0.131	19549.995
$\boldsymbol{\beta}_{11}$: 2×5	0.021	0.021	-0.046	0.089	19457.065
$\boldsymbol{\beta}_{12}$: 3×4	0.095	0.095	0.046	0.145	25369.488
$\boldsymbol{\beta}_{13}$: 3×5	0.193	0.193	0.145	0.243	23725.779
$\boldsymbol{\beta}_{14}$: 4×5	-0.051	-0.051	-0.100	-0.002	24857.852
$\boldsymbol{\beta}_{15}$: $1 \times 3 \times 4$	-0.037	-0.037	-0.102	0.031	18689.917
$\boldsymbol{\beta}_{16}$: $2 \times 3 \times 4$	0.046	0.046	-0.021	0.113	18811.420
$\boldsymbol{\beta}_{17}$: $1 \times 3 \times 5$	-0.065	-0.065	-0.132	0.002	18661.653
$\boldsymbol{\beta}_{18}$: $2 \times 3 \times 5$	0.161	0.161	0.094	0.230	19234.499
$\boldsymbol{\beta}_{19}$: $1 \times 4 \times 5$	-0.042	-0.042	-0.110	0.023	19796.697
$\boldsymbol{\beta}_{20}$: $2 \times 4 \times 5$	-0.004	-0.004	-0.071	0.063	20054.320
$\boldsymbol{\beta}_{21}$: $3 \times 4 \times 5$	-0.012	-0.012	-0.061	0.037	25091.911
$\boldsymbol{\beta}_{22}$: $1 \times 3 \times 4 \times 5$	0.039	0.039	-0.026	0.107	19263.891
$\boldsymbol{\beta}_{23}$: $2 \times 3 \times 4 \times 5$	-0.050	-0.050	-0.118	0.017	18554.812
σ_s	0.366	0.362	0.276	0.462	11770.313

Note: The effects coded variables were as follows: (1) Shape = 1, Binding = 0, Colour = -1, (2) Shape = 0, Binding = 1, Colour = -1, (3) SS3 = -1, SS6 = 1, (4) Younger = -1, Older = 1, (5) No-Change = -1, Change = 1. Interaction contrasts were products of these effects coded variables.

type (colour change, shape change, binding change, no-change), set size (3, 6), and age-group (younger, older).

Specific hypothesis tests show that accuracy was lower for 6 items than for 3, -0.686 [-0.815, -0.555], and for older adults relative to younger adults, -0.600 [-0.767, -0.436]. However, set size did not clearly modulate age-differences in performance, 0.079 [-0.188, 0.336]. Further, comparing the colour and shape change trials to trials on which there was a binding change we find that feature change detection was far more accurate than detection of a binding change, 0.837 [0.678, 0.989]. Crucially

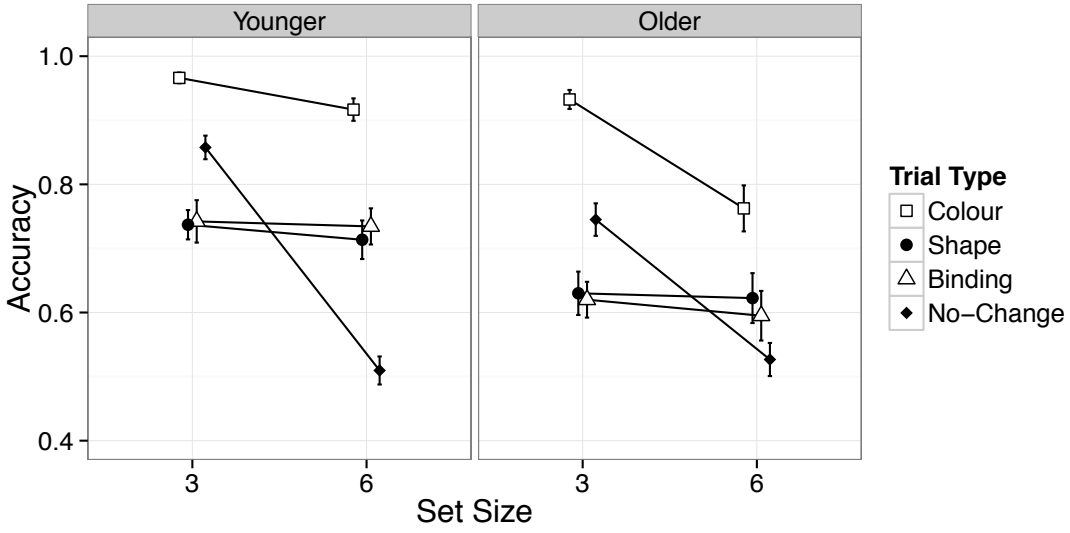

Figure 4.3: Accuracy for mixed trials across age-groups and experimental conditions. Error bars are ± standard error.

the discrepancy between feature and binding performance was similar across the age-groups, -0.120 [-0.425, 0.196]. Once again (see above section), although the HDI overlaps zero, the direction of this contrast signifies that the binding cost was if anything smaller in the older group. Increasing the number of memory items had no discernible effect on the age by change type (feature versus binding) interaction, -0.143 [-0.769, 0.487]. Finally, the slight age-group by trial type (change versus no-change) interaction was also present in the mixed data set. The effect of age on accuracy was less pronounced for the no-change trials in contrast with (the average of) the change trials, -0.179 [-0.285, -0.073].

Mixed Versus Blocked Analysing the raw accuracy for the mixed and blocked participants separately is informative in showing us the overall patterns of responding across these conditions but in order to pin down differences between the two we must combine the data sets and analyse as one. As no-change trials in the mixed condition cannot be distinguished for colour, shape and binding this analysis focuses on accuracy for change trials only.

Table 4.4 presents the results of a logit model with factors of trial type (colour change, shape change, binding change), set size (3, 6), age-group (younger, older), and block type (mixed, blocked). Interestingly participants were better at detecting

Table 4.3: Posterior quantities from logit model for the Mixed condition

			95% HDI		
Parameter	Mean	Median	lower	upper	ESS
β_0	1.167	1.167	1.082	1.251	6140.908
$\boldsymbol{\beta}_1$: (1) Shape	-0.416	-0.416	-0.516	-0.315	20718.657
$\boldsymbol{\beta}_2$: (2) Binding	-0.422	-0.422	-0.523	-0.325	21137.377
$\boldsymbol{\beta}_3$: (3) No-Change	-0.408	-0.408	-0.490	-0.330	10590.778
$\boldsymbol{\beta}_4$: (4) SS6	-0.343	-0.343	-0.408	-0.278	9492.274
$\boldsymbol{\beta}_5$: (5) Older Group	-0.300	-0.300	-0.383	-0.218	6656.232
$\boldsymbol{\beta}_6$: 1×4	0.305	0.305	0.208	0.406	19677.802
$\boldsymbol{\beta}_7$: 2×4	0.306	0.306	0.210	0.407	20769.862
$\boldsymbol{\beta}_8$: 3×4	-0.343	-0.343	-0.424	-0.263	10287.139
$\boldsymbol{\beta}_9$: 1×5	0.069	0.069	-0.031	0.167	20621.771
$\boldsymbol{\beta}_{10}$: 2×5	-0.005	-0.005	-0.103	0.095	18681.638
$\boldsymbol{\beta}_{11}$: 3×5	0.134	0.134	0.055	0.214	11079.891
$\boldsymbol{\beta}_{12}$: 4×5	0.020	0.020	-0.047	0.084	9411.368
$\boldsymbol{\beta}_{13}$: $1 \times 4 \times 5$	0.002	0.003	-0.095	0.106	20324.215
$\boldsymbol{\beta}_{14}$: $2 \times 4 \times 5$	-0.036	-0.036	-0.134	0.066	20673.431
$\boldsymbol{\beta}_{15}$: $3 \times 4 \times 5$	0.180	0.180	0.100	0.261	10208.954
σ_s	0.184	0.183	0.113	0.260	3565.263

Note: The effects coded variables were as follows: (1) Shape = 1, Binding = 0, No-Change = 0, Colour = -1, (2) Shape = 0, Binding = 1, No-Change = 0, Colour = -1, (3) Shape = 0, Binding = 0, No-Change = 1, Colour = -1, (4) SS3 = -1, SS6 = 1, (5) Younger = -1, Older = 1. Interaction contrasts were products of these effects coded variables.

changes in the mixed condition relative to the blocked condition, 0.334 [0.121, 0.555], and this difference between block types was similar across the age-groups, -0.043 [-0.483, 0.383]. Change detection was better when a change had occurred to an individual feature (colour or shape only) relative to a feature swap, 0.559 [0.451, 0.665]. There was no clear effect of age on the discrepancy between feature and binding change detection, -0.107 [-0.324, 0.109]. However, of primary interest is whether or not mixing conjunction changes with changes to individual features has a disproportionate effect on healthy older adults' ability to detect those changes. Our data suggest that this is not the case; age differences in the binding cost were no larger with mixed relative to blocked trials, -0.016 [-0.437, 0.425].

Our analysis of raw accuracy is in line with the findings of T. Chen and Naveh-Benjamin (2012). Older adults were less likely to detect changes but the ability to

Table 4.4: Posterior quantities from logit model comparing change trials in the mixed and blocked conditions

			95% HDI		
Parameter	Mean	Median	lower	upper	ESS
β_0	1.179	1.180	1.070	1.288	3552.891
$\boldsymbol{\beta}_1$: (1) Shape	-0.510	-0.510	-0.579	-0.438	18647.387
$\boldsymbol{\beta}_2$: (2) Binding	-0.373	-0.373	-0.444	-0.301	18406.937
$\boldsymbol{\beta}_3$: (3) SS6	-0.215	-0.214	-0.271	-0.160	15768.569
$\boldsymbol{\beta}_4$: (4) Older Group	-0.336	-0.336	-0.441	-0.224	3574.116
$\boldsymbol{\beta}_5$: (5) Mixed	0.167	0.167	0.061	0.277	3579.921
$\boldsymbol{\beta}_6$: 1×3	0.110	0.110	0.039	0.182	17853.982
$\boldsymbol{\beta}_7$: 2×3	0.225	0.225	0.155	0.298	18700.585
$\boldsymbol{\beta}_8$: 1×4	0.080	0.080	0.009	0.152	17922.860
$\boldsymbol{\beta}_9$: 2×4	0.036	0.036	-0.036	0.108	19495.189
$\boldsymbol{\beta}_{10}$: 1×5	-0.055	-0.055	-0.123	0.018	17230.747
$\boldsymbol{\beta}_{11}$: 2×5	-0.199	-0.198	-0.270	-0.127	18597.355
$\boldsymbol{\beta}_{12}$: 3×4	0.021	0.021	-0.036	0.076	15911.808
$\boldsymbol{\beta}_{13}$: 3×5	-0.018	-0.018	-0.074	0.036	15803.075
$\boldsymbol{\beta}_{14}$: 4×5	-0.011	-0.011	-0.121	0.096	3583.536
$\boldsymbol{\beta}_{15}$: $1 \times 3 \times 4$	0.033	0.033	-0.035	0.106	18225.172
$\boldsymbol{\beta}_{16}$: $2 \times 3 \times 4$	0.011	0.011	-0.062	0.081	17924.071
$\boldsymbol{\beta}_{17}$: $1 \times 3 \times 5$	0.083	0.083	0.012	0.153	19307.092
$\boldsymbol{\beta}_{18}$: $2 \times 3 \times 5$	-0.030	-0.030	-0.101	0.041	17477.389
$\boldsymbol{\beta}_{19}$: $1 \times 4 \times 5$	0.034	0.034	-0.039	0.102	18294.783
$\boldsymbol{\beta}_{20}$: $2 \times 4 \times 5$	0.003	0.003	-0.071	0.073	17578.808
$\boldsymbol{\beta}_{21}$: $3 \times 4 \times 5$	-0.063	-0.063	-0.119	-0.007	16426.652
$\boldsymbol{\beta}_{22}$: $1 \times 3 \times 4 \times 5$	0.031	0.031	-0.042	0.097	18530.034
$\boldsymbol{\beta}_{23}$: $2 \times 3 \times 4 \times 5$	0.015	0.015	-0.057	0.088	17704.231
σ_s	0.463	0.461	0.375	0.555	10094.882

Note: The effects coded variables were as follows: (1) Shape = 1, Binding = 0, Colour = -1, (2) Shape = 0, Binding = 1, Colour = -1, (3) SS3 = -1, SS6 = 1, (4) Younger = -1, Older = 1, (5) Blocked = -1, Mixed = 1. Interaction contrasts were products of these effects coded variables.

detect changes to colour-shape binding did not exhibit a disproportionate effect of age. This was true whether or not feature and binding changes were encountered together in the same trial-block or separately in their own blocks of trials. However, it is possible that important differences between our two age-groups may have been obscured in the analysis of raw accuracy. Therefore, in a subsequent analysis we attempted to separate out contributions of sensitivity and response bias to performance on this task. Cowan et al. (2006) found that older adults employed a

strict response criterion for binding changes as well as being less sensitive to these changes when they were mixed with colour changes. This analysis also allows us to use Rouder et al. (2012)'s default Bayes factors to assess the weight of evidence *against* age × condition interactions.

Sensitivity and Bias

In their analysis of performance Cowan et al. (2006) attempted to separate the contribution of sensitivity and bias to younger and older adults' performance on their change detection tasks using the signal detection theory metrics, d' for sensitivity and c for criterion placement. When colour and colour-location change trials were mixed within the same block older adults were less sensitive to binding changes as well as showing a bias towards responding 'same'.

Here we also attempt to separate out these two contributions to the raw accuracy data reported above. Subsequent work has assessed the receiver operating characteristics (see Chapter 8) of the standard VWM change detection task and has suggested that the process underlying performance on this task is more consistent with a threshold process rather than comparison of a random variable to a criterion (Donkin et al., 2013; Rouder et al., 2008). Thus we report the two-high threshold measures P_r for sensitivity and B_r for response bias (Snodgrass & Corwin, 1988) in our main analysis. Analysing d' and c does not change our main conclusions but, as Chapter 8 highlights, this is certainly not guaranteed and these measures should not be used interchangeably as they imply different models of the recognition process.

Sensitivity Figure 4.4 presents the sensitivity measure, P_r, for both the blocked and mixed conditions across age-groups. Comparing the left and right panels we see surprisingly little difference between the pattern of performance between the two age groups, beyond generally lower performance in the older group. Tables 4.5 and 4.6 present the results of Bayesian ANOVAs on sensitivity (P_r) for the blocked and mixed conditions, respectively.

For the blocked condition the 'winning' model was made up of main effects of memory condition, set size, and age-group with no interactions. Comparing this

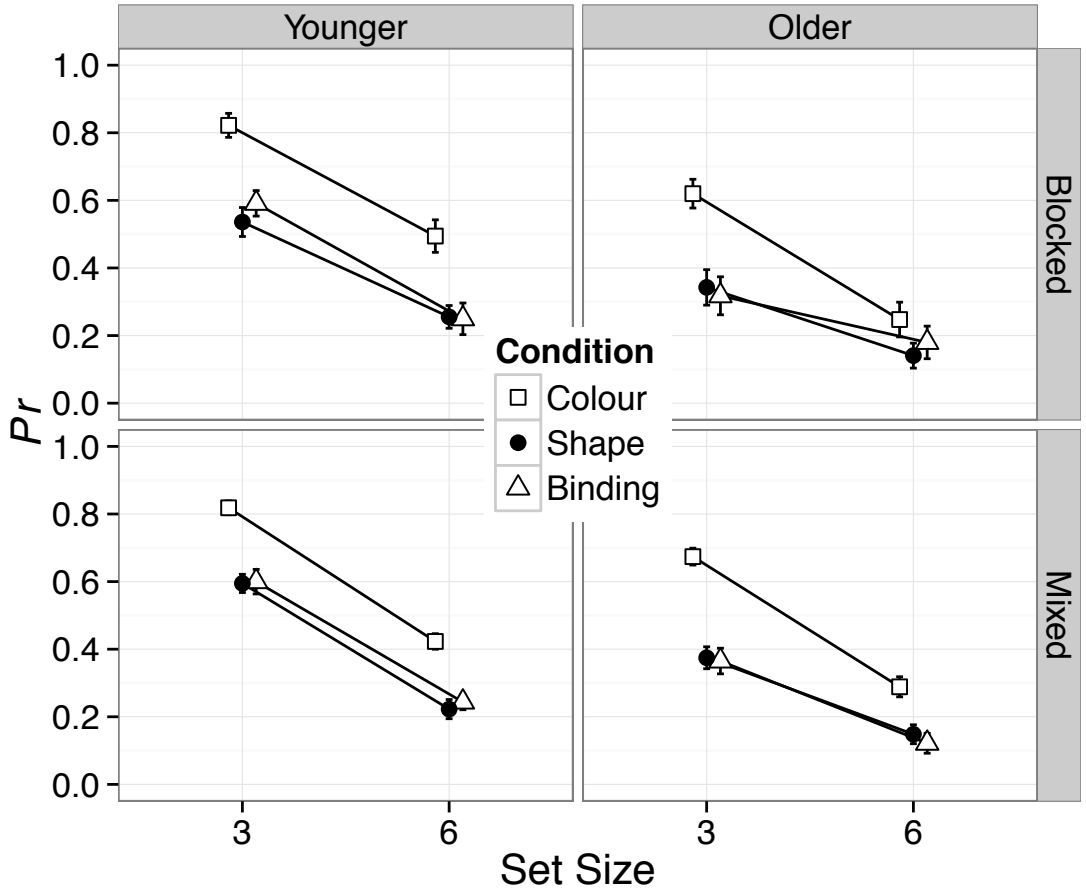

Figure 4.4: P_r (corrected recognition) across age groups and experimental conditions. Error bars are ± standard error.

model 1 to the sixth model we find 6 to 1 evidence against a memory condition by age-group interaction. There was rather weak evidence in favour of the three-way interaction between age, memory condition and set size ($B_{5,9} = 1.22$). This was clearly due to a smaller set size effect for the older group in the shape and binding conditions (see top panels of Figure 4.4). In the mixed condition the winning model included all main effects as well as age-group × set size and memory condition × set size interactions. As with the analysis of the blocked condition, the weight of evidence was against the crucial two-way interaction between age and condition ($B_{1,3} = 8.25$). For this analysis there was also weak evidence against the three-way interaction ($B_{3,4} = 1.68$).

In addition to the analyses presented in Tables 4.5 and 4.6 a Bayesian ANOVA was conducted on the full data set with the additional factor of block type (mixed

Table 4.5: Log Bayes factors for analysis of sensitivity (P_r) in the blocked condition

	Model	$\log(B_{M,0})$	% error
1	$P_r \sim$ C + SS + AG + ID	86.86	1.23
2	$P_r \sim$ C + SS + AG + SS:AG + ID	86.72	0.95
3	$P_r \sim$ C + SS + C:SS + AG + ID	86.69	1.09
4	$P_r \sim$ C + SS + C:SS + AG + SS:AG + ID	86.62	1.19
5	$P_r \sim$ C + SS + C:SS + AG + C:AG + SS:AG + C:SS:AG + ID	85.00	1.99
6	$P_r \sim$ C + SS + AG + C:AG + ID	85.00	1.14
7	$P_r \sim$ C + SS + AG + C:AG + SS:AG + ID	84.88	1.27
8	$P_r \sim$ C + SS + C:SS + AG + C:AG + ID	84.88	1.48
9	$P_r \sim$ C + SS + C:SS + AG + C:AG + SS:AG + ID	84.80	1.27
10	$P_r \sim$ C + SS + ID	81.55	0.62
11	$P_r \sim$ C + SS + C:SS + ID	81.46	4.25
12	$P_r \sim$ SS + AG + ID	51.79	0.49
13	$P_r \sim$ SS + AG + SS:AG + ID	51.23	0.91
14	$P_r \sim$ SS + ID	46.63	0.39
15	$P_r \sim$ C + AG + ID	25.52	0.53
16	$P_r \sim$ C + AG + C:AG + ID	23.35	1.18
17	$P_r \sim$ C + ID	20.54	0.90
18	$P_r \sim$ AG + ID	4.74	0.23

Note: All Bayes factors are relative to a null model with a random participant effect only (model 0: $P_r \sim$ ID). AG = Age-Group, C = Condition, SS = Set Size, ID = participant ID, and ':' denotes an interaction effect

or blocked). This analysis compared a full model to reduced models omitting a single component (main- or interaction-effect) at a time, thus reducing the number of models to be computed (see Brown et al., 2016, for the same approach). When omitting the age × memory condition interaction the resulting reduced model was approximately 15 times more likely, given the data, than the full model. This large data set, therefore, provides the strongest evidence observed so far *against* a disproportionate effect of age in a particular condition. As is typical in studies like this one, we conducted an additional analysis omitting the colour condition to contrast binding with the most difficult individual feature condition, shape. This analysis also favoured the omission of the interaction ($B_{R,F} = 4.902$).

An analysis of d' revealed strong evidence *for* the interaction between age and condition ($B_{R,F} = 0.101$, around 10-to-1 in favour of the interaction). This was clearly due to a smaller effect of age in the colour condition as an analysis of the shape and binding conditions only, favoured omission of the interaction ($B_{R,F} = 2.043$). Nevertheless, for the reasons discussed above, we favour the analysis of P_r.

Omitting the three-way interaction between age, condition, and block type resulted in a model that was favoured over the full model by a factor of 4.7-to-1 (for

Table 4.6: Log Bayes factors for analysis of sensitivity (P_r) in the mixed condition

	Model	$\log(B_{M,0})$	% error
1	$P_r \sim$ C + SS + C:SS + AG + SS:AG + ID	172.22	1.49
2	$P_r \sim$ C + SS + AG + SS:AG + ID	170.95	0.87
3	$P_r \sim$ C + SS + C:SS + AG + C:AG + SS:AG + ID	170.11	1.39
4	$P_r \sim$ C + SS + C:SS + AG + C:AG + SS:AG + C:SS:AG + ID	169.59	3.09
5	$P_r \sim$ C + SS + C:SS + AG + ID	169.46	3.69
6	$P_r \sim$ C + SS + AG + C:AG + SS:AG + ID	168.83	1.03
7	$P_r \sim$ C + SS + AG + ID	168.28	0.50
8	$P_r \sim$ C + SS + C:SS + AG + C:AG + ID	167.27	1.45
9	$P_r \sim$ C + SS + AG + C:AG + ID	166.14	0.68
10	$P_r \sim$ C + SS + C:SS + ID	158.81	0.62
11	$P_r \sim$ C + SS + ID	157.71	0.46
12	$P_r \sim$ SS + AG + SS:AG + ID	102.30	2.52
13	$P_r \sim$ SS + AG + ID	101.53	0.63
14	$P_r \sim$ SS + ID	91.86	0.42
15	$P_r \sim$ C + AG + ID	32.08	1.30
16	$P_r \sim$ C + AG + C:AG + ID	29.57	1.16
17	$P_r \sim$ C + ID	24.53	0.19
18	$P_r \sim$ AG + ID	6.38	0.26

Note: All Bayes factors are relative to a null model with a random participant effect only (model 0: $P_r \sim$ ID). AG = Age-Group, C = Condition, SS = Set Size, ID = participant ID, and ':' denotes an interaction effect

the shape and binding conditions: $B_{R,F} = 4.684$). It seems, then, that mixing individual feature and binding trials within the same block had no effect on older adults' sensitivity to change across the different memory conditions. Further, there was no suggestion of a four-way interaction in either the full analysis ($B_{R,F} = 3.985$) or the restricted analysis of the shape and binding conditions ($B_{R,F} = 2.047$).

Beyond the interactions of primary interest to the present work, there was strong evidence for the interaction between age and set size ($B_{F,R} = 34.008$). There was a much larger effect of age at set size 3 relative to set size 6; this was unexpected however it is possible that at a lower set size younger adults were more able to employ elaborative strategies to boost performance (R. J. Allen et al., 2006; Mitchell, Johnson, Raye, Mather, & D'Esposito, 2000) but when presented with 6 items were less able to do this and thus exhibited a greater performance cost. Another potential account is that older adults' performance with 6 to-be-remembered items had hit an effective floor, although we note that performance was always above chance level. In addition there was slight evidence for a three-way interaction between age, memory condition, and set size ($B_{F,R} = 3.25$). As Figure 4.4 shows, for the older adults the effect of increasing the number of items was less pronounced for shape and

binding conditions. This, again, may suggest that older adults' performance was at an effective floor, and that consequently we were less able to find a disproportionate binding deficit. While this cannot be ruled out, Experiments 7 and 8 reported in the next Chapter also found no-evidence for an age-related binding deficit with much higher levels of performance.

Finally mixing different trial types together did not modulate the effect of age on change detection sensitivity, as a model omitting the age by block type interaction was favoured over the full model by a factor of 3-to-1. This is particularly intriguing given the potential the comparison between mixed and blocked conditions has to reveal potential strategic differences between younger and older adults. Indeed, mixing trials appeared to have no effect on performance at all as omitting the main effect of block type was also favoured ($B_{R,F} = 5.186$). The ability to prepare for a specific type of change in the blocked condition did not appear to benefit performance, suggesting that the change detection task may be rather unaccommodating of strategy use (C. C. Morey & Cowan, 2004).

Bias The guessing bias measure, B_r, across conditions and groups is presented in Figure 4.5. In general our older adults exhibit less extreme bias—their data points tend to lie closer to the neutral 0.5 level. However, as was the case for sensitivity, the overall pattern is remarkably similar between the two groups (compare left and right panels of Figure 4.5).

The results of a Bayesian ANOVA on guessing bias (B_r), in the blocked condition are presented in Table 4.7. The winning model contains main effects of memory condition and set size as well as their interaction. That this model was preferred over model 2, which also includes the effect of age-group, provides evidence that age had no overall effect on bias in this condition ($B_{1,2} = 2.649$). Further, there was strong evidence against the crucial interaction between memory condition and age ($B_{2,4} = 11.3$) and good evidence against modulation by set size ($B_{5,8} = 4.964$). In summary the pattern seen in the top panels of Figure 4.5 is supported by our BANOVA; response bias did not vary greatly across age groups.

For the analysis of the mixed data the winning model contained main effects

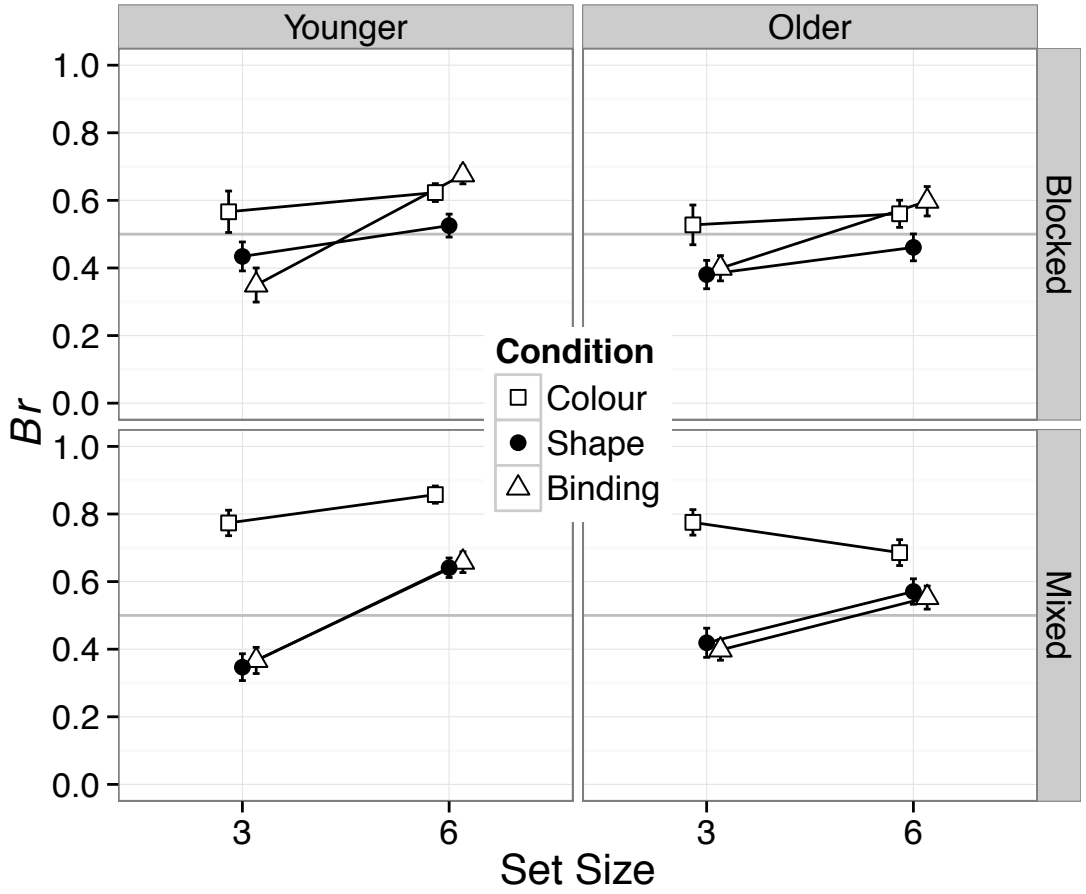

Figure 4.5: B_r across age groups and experimental conditions. Error bars are ± standard error.

of age-group, memory condition and set size with two-way interactions between set size and condition as well as set size and age group. Comparing models 1 and 5 we see that there was overwhelming evidence for the age by set size interaction ($B_{1,5}$ = 5611.209). It is clear from Figure 4.5 that younger adults tended to go from a preference towards guessing 'no-change' at set size 3 towards preferring 'change' at set size 6. This tendency was far less pronounced for the older group (see the bottom panels of Figure 4.5).

Crucially, as in the blocked analysis, the winning model did not include the interaction between age and condition. However, comparing models 1 and 2 we find only a marginal preference for the model omitting this interaction ($B_{1,2}$ = 1.782). Again, scrutinising Figure 4.5 shows clearly that older adults did not exhibit extreme bias in the binding condition. Finally there was good evidence against the three-way

Table 4.7: Log Bayes factors for analysis of guessing bias (B_r) in the blocked condition

	Model	$\log(B_{M,0})$	% error
1	$B_r \sim$ C + SS + C:SS + ID	20.59	0.69
2	$B_r \sim$ C + SS + C:SS + AG + ID	19.62	3.37
3	$B_r \sim$ C + SS + C:SS + AG + SS:AG + ID	18.51	0.98
4	$B_r \sim$ C + SS + C:SS + AG + C:AG + ID	17.19	0.78
5	$B_r \sim$ C + SS + C:SS + AG + C:AG + SS:AG + ID	16.14	1.36
6	$B_r \sim$ C + SS + ID	15.94	0.62
7	$B_r \sim$ C + SS + AG + ID	14.90	0.71
8	$B_r \sim$ C + SS + C:SS + AG + C:AG + SS:AG + C:SS:AG + ID	14.54	4.81
9	$B_r \sim$ C + SS + AG + SS:AG + ID	13.79	0.99
10	$B_r \sim$ C + SS + AG + C:AG + ID	12.50	0.79
11	$B_r \sim$ SS + ID	11.48	0.28
12	$B_r \sim$ C + SS + AG + C:AG + SS:AG + ID	11.40	1.12
13	$B_r \sim$ SS + AG + ID	10.42	1.07
14	$B_r \sim$ SS + AG + SS:AG + ID	9.26	0.78
15	$B_r \sim$ C + ID	3.67	0.27
16	$B_r \sim$ C + AG + ID	2.57	0.33
17	$B_r \sim$ C + AG + C:AG + ID	0.15	0.76
18	$B_r \sim$ AG + ID	-1.11	1.27

Note: All Bayes factors are relative to a null model with a random participant effect only (model 0: $B_r \sim$ ID). AG = Age-Group, C = Condition, SS = Set Size, ID = participant ID, and ':' denotes an interaction effect

interaction in this condition ($B_{2,3} = 7.825$).

Turning to the analysis of the complete data set, with the additional factor of block type, this revealed that a model omitting the main effect of age was very marginally preferred over the full model ($B_{R,F} = 1.678$). Thus there was no discernible overall effect of age on response bias. The model omitting the age by memory condition interaction was preferred over the full model by more than 11-to-1. Therefore, in addition there being no clear effect of age on response bias it seems that this was true regardless of the to-be-remembered features. Further, the type of block did not modulate this, as a model omitting the three-way interaction was favoured ($B_{R,F} = 5.24$). The weight of evidence was against all of the other interactions including age-group with the exception of the interaction between age and set size. The full model was favoured by 96-to-1 over the model omitting the interaction. As discussed above, increasing the size of the memory array had a greater effect on the guessing bias exhibited by our younger adults (see Figure 4.5). In summary, our two age-groups do not appear to differ greatly in response bias and when differences do arise younger adults tended to exhibit more extreme guessing

Table 4.8: Log Bayes factors for analysis of guessing bias (B_r) in the mixed condition

	Model	$\log(B_{M,0})$	% error
1	$B_r \sim$ C + SS + C:SS + AG + SS:AG + ID	110.60	1.58
2	$B_r \sim$ C + SS + C:SS + AG + C:AG + SS:AG + ID	110.02	2.41
3	$B_r \sim$ C + SS + C:SS + AG + C:AG + SS:AG + C:SS:AG + ID	107.97	1.01
4	$B_r \sim$ C + SS + C:SS + ID	102.79	0.55
5	$B_r \sim$ C + SS + C:SS + AG + ID	101.97	0.91
6	$B_r \sim$ C + SS + C:SS + AG + C:AG + ID	101.23	0.80
7	$B_r \sim$ C + SS + AG + SS:AG + ID	93.94	1.11
8	$B_r \sim$ C + SS + AG + C:AG + SS:AG + ID	93.04	0.61
9	$B_r \sim$ C + SS + ID	87.76	0.44
10	$B_r \sim$ C + SS + AG + ID	86.87	0.54
11	$B_r \sim$ C + SS + AG + C:AG + ID	85.85	0.58
12	$B_r \sim$ C + ID	60.44	0.21
13	$B_r \sim$ C + AG + ID	59.49	0.99
14	$B_r \sim$ C + AG + C:AG + ID	58.10	0.64
15	$B_r \sim$ SS + AG + SS:AG + ID	16.31	1.42
16	$B_r \sim$ SS + ID	14.58	1.19
17	$B_r \sim$ SS + AG + ID	13.42	0.35
18	$B_r \sim$ AG + ID	-1.18	0.60

Note: All Bayes factors are relative to a null model with a random participant effect only (model 0: $B_r \sim$ ID). AG = Age-Group, C = Condition, SS = Set Size, ID = participant ID, and ':' denotes an interaction effect

biases.

4.3 General Discussion

In summary our analyses, both of raw accuracy and indices of recognition sensitivity and bias, provide no evidence to suggest that older adults specifically struggle to detect binding changes. Indeed, using default Bayes factors (Rouder et al., 2012), we were able to provide evidence against this suggestion for both sensitivity and bias (see also Brown et al., 2016; S. Rhodes et al., 2016). For sensitivity the data favoured the absence of a differential effect of age across the memory conditions by over 15-to-1 and for bias 11-to-1. Further the weight of evidence was against any effect of block type. Rather there appears to be a more general decline of VWM recognition performance with age (see also, for example, W. Johnson et al., 2010; Sander et al., 2011a) with a particular effect on the ability to detect changes as opposed to no-change. We attempt to pick apart potential contributory factors to this generally poorer recognition in a subsequent exploratory modelling section (Chapter 7).

Taking the present findings with other research examining similar questions (Brockmole et al., 2008; Brockmole & Logie, 2013; Brown et al., 2016; Isella et al., 2015; S. Rhodes et al., 2016) and our reanalysis of Brown and Brockmole (2010) in Chapter 3 it is becoming clear that healthy ageing does not disproportionately affect the ability to bind the surface features (i.e. shape and colour) of objects and retain these conjunctions in VWM. The present work builds on this by showing that mixing feature and binding change trials within the same test block had no discernible effect on older adults' recognition performance. This is in-line with the recent findings of T. Chen and Naveh-Benjamin (2012) who were assessing associative recognition memory but contrary to the results of Cowan et al. (2006) who used a paradigm far more similar to our own.

Nevertheless, there are a number of differences between the present study and the paradigm of Cowan et al. (2006) that may account for the discrepant findings. One possibility is that participants in the experiments of Cowan et al. (2006) were not motivated to attend to each feature (colour, location) equally. In their experiments only two kinds of change were included; one in which the circled probe item was a brand new colour (item change) and one where the probe was a brand new colour-location pairing (binding change). In the mixed condition knowledge that around half of the changes would occur to colour only may have induced a strategic bias towards focusing attention on colours present in the array at the expense of their precise location (Woodman & Vogel, 2008). In our study, as well as the study of T. Chen and Naveh-Benjamin (2012), we included changes to *both* individual features in addition to the possibility of conjunction changes. In the following Chapter we directly assess the effect of including only one kind of feature change in the context of colour-location binding and find no evidence that this has any effect on change detection performance.

Another potentially important methodological difference between the present work and that of Cowan et al. (2006) is the nature of our probe array. In both cases participants made a judgement on a single probed item, however, in Cowan et al. (2006) unprobed items were also present in the test array. Further, as Cowan et al.

(2006) used arrays as large as ten items, their stimuli were selected *with* replacement from a set of seven colours, so both memory and test arrays could contain repeated colours. The presence of duplicates meant that for colour trials the probed item was always a unique colour in the array, whereas for binding trials the probed item was always repeated. As noted by Cowan et al. (2014) for colour trials it would have been sufficient for observers to retain the unique colours from the array, whereas for binding trials it would have been enough to notice which colours were duplicated. This aspect of the task, if noted by participants, may have altered relative change detection performance between the feature and binding change trials. Thus the single probe methodology used here is arguably better for assessing the evidence for age-related binding deficits.

The experiments of Cowan et al. (2006) assessed the lifespan development of binding features (colours) to spatial location. It is possible that the age-related binding deficit found in their first experiment arose due to the requirement to remember *what was where*. Indeed many of the reports suggesting that older adults have a specific difficulty binding information in VWM have used location as a critical feature (Borg et al., 2011; Fandakova et al., 2014; Mitchell, Johnson, Raye, Mather, & D'Esposito, 2000; Mitchell, Johnson, Raye, & D'Esposito, 2000). Given the role of the hippocampus in allocentric spatial processing (e.g., Ekstrom et al., 2003; O'Keefe & Nadel, 1978) and the pronounced volumetric and functional changes to this region with age (e.g., Raz & Rodrigue, 2006; Mitchell, Johnson, Raye, & D'Esposito, 2000) this has led to the suggestion that older adults *specifically* struggle to bind to location in WM (Brockmole et al., 2008). Recent work has cast doubt on this suggestion (Pertzov et al., 2015; Read et al., 2016) and in the next Chapter we provide a critical summary of the existing evidence and find that it is far less compelling than previously thought. Nevertheless, the binding of surface features as assessed in the present study has not, to our knowledge, been compared to location binding using comparable experimental procedures. Therefore, we conducted an additional set of experiments assessing younger and older adults' ability to temporarily retain the correspondence between colour and location in WM.

To summarise, the present experiment further assessed the ability of healthy older adults to retain combinations of surface features over a brief interval. In line with a growing literature, the effect of age was no greater in conditions requiring the detection of binding changes compared to conditions requiring the detection of feature changes. Further, we obtained clear evidence that mixing different kinds of change to object features *does not* influence younger and older adults sensitivity to these changes, in contrast to previous work. While there are a number of methodological differences between the present work and previous studies we can make a strong case that the procedure use here is better placed to address our key questions. A large number of studies under various experimental conditions have failed to demonstrate a specific age-related colour-shape binding deficit. However, the present literature suggests that this may not be the case for all forms of VWM binding. To build on these findings in Chapter 5 we go on to assess colour-location binding to address the commonly made suggestion that older adults have a specific problem binding to location in VWM.

Chapter 5

Ageing and Feature Binding: Is Location Special?

5.1 Introduction

The previous Chapter assessed the effect of mixing feature and conjunction changes within the same block of trials on older adults' ability to detect changes to colour-shape binding. This was motivated by the findings of Cowan et al. (2006) who found that older adults were less sensitive to colour-location binding changes when they were mixed in with changes to colour only relative to when these were presented in separate trial-blocks. There are a number of methodological differences between these two studies, however a crucial one may be that we assessed binding between surface features whereas Cowan et al. (2006) assessed binding between surface and spatial features. Early studies by Mitchell and colleagues (Mitchell, Johnson, Raye, Mather, & D'Esposito, 2000; Mitchell, Johnson, Raye, & D'Esposito, 2000) claimed to show that the effect of age was far larger for tasks requiring participants to retain the binding of object and location in working memory relative to tasks requiring the retention of the component parts alone. Subsequent studies have made similar claims that older adults struggle to retain *what was where* in WM (Borg et al., 2011; Fandakova et al., 2014; Peich et al., 2013). On the other hand studies assessing binding between object's colour and shape have largely found no evidence for an age-

related binding deficit (Brockmole et al., 2008; Brockmole & Logie, 2013; Brown et al., 2016; Isella et al., 2015; S. Rhodes et al., 2016). Therefore, it may be that we found no role for mixing versus blocking trial types as older adults' ability to retain surface feature bindings is relatively intact. The link between identity and location, however, may become more fragile with age and thus may explain the discrepancy between our findings in Experiment 6 (Chapter 4) and the findings of Cowan and colleagues.

Consequently we conducted an additional set of experiments assessing older adults' ability to retain colour-location conjunctions using a similar methodology to our previous experiment. Feature binding between surface features has not, to our knowledge, been compared to surface-spatial binding using comparable paradigms before in the context of healthy ageing and this should provide a good test of suggestions that binding to location is a specific problem for older adults (see, Brockmole et al., 2008). Before outlining this study, however, it is important to evaluate the quality of evidence for an age-related deficit in retaining what was where over brief periods.

The evidence for location binding deficits with age

Mitchell and colleagues were the first to assess the effect of age on the retention of feature bindings in working memory in a series of experiments published over 15 years ago (Mitchell, Johnson, Raye, Mather, & D'Esposito, 2000; Mitchell, Johnson, Raye, & D'Esposito, 2000). In two behavioural experiments they presented younger and older participants (24 of each in Experiment 1 and 16 in Experiment 2) with a sequence of three nameable clip-art-like objects (for 1 second each) on a 3×3 grid and then probed memory for the object, location, or object-location pairing following an 8.5 second delay (Mitchell, Johnson, Raye, Mather, & D'Esposito, 2000). In both experiments t-tests revealed that older adults' recognition performance (in terms of d') was significantly poorer than that of younger adults but only in blocks requiring the retention of object-location binding. There was no significant performance difference when the task required that individual features were held in WM.

This pattern was also found in a functional neuroimaging study using a similar task with much smaller groups (6 in each) of younger and older adults and, this time, corrected recognition (P_r) as the measure of performance (Mitchell, Johnson, Raye, & D'Esposito, 2000). Although we note that the authors applied a rather liberal criterion for the age-group difference in the binding condition ($p = 0.06$).

Mitchell et al. argued that, as there was a significant group difference in the condition requiring participants to detect changes to object-location conjunctions but not in conditions requiring them to detect changes to features only, there was a specific age-related deficit in forming and retaining feature bindings in WM. Unfortunately, however, this is not true and is an expression of a statistical fallacy (see, Gelman & Stern, 2006; Nieuwenhuis, Forstmann, & Wagenmakers, 2011). Finding a significant age-group difference in one condition but not another is not sufficient to support the suggestion that the effect of age is specifically pronounced in one condition relative to another. What this requires is the evaluation of the age-group $\times$ condition interaction. Crucially, in the behavioural experiments of Mitchell, Johnson, Raye, Mather, and D'Esposito (2000) this interaction was *not* significant by conventional standards ($p = 0.06$ and 0.13 for Experiments 1 and 2, respectively). In the fMRI study reported by Mitchell, Johnson, Raye, and D'Esposito (2000) the test of the interaction is not even reported, however, given that the difference between the age-groups in the object-location conjunction condition was not significant at conventional levels (as noted above) we can be fairly sure that it is not significant in this study either. Thus while the findings of Mitchell and colleagues have been taken to indicate a specific age-related deficit in retaining object-location correspondences further scrutiny reveals that the evidence is far from convincing.

This basic statistical error appears repeatedly in the literature on binding deficits in VWM. Borg et al. (2011) reported a study assessing object-location binding in groups of healthy younger and older adults, as well as a group of patients with mild-AD (14 in each group). In one condition participants had to remember pictures depicting neutral or negatively valenced objects only, whereas in another condition they had to remember the correspondence between the picture and its presented

location. The analysis presented in this paper proceeds in a confusing fashion; the authors conducted a one-way ANOVA on (arcsin transformed) accuracy comparing the two conditions (pictures only versus picture-location conjunctions), finding better overall performance in the pictures only condition. This was followed up by *separate* two-way ANOVAs for each condition with group and valence as independent variables. The crucial finding was that in the picture only condition there was no group by valence interaction, whereas this interaction was significant for the binding condition. It seemed, from post-hoc testing, that healthy older adults performed better than the Alzheimer's group for neutral images but not for negative ones. However, it is not clear from the analysis presented by Borg et al. (2011) whether this pattern is specific to the binding condition (which would have required assessment of the three-way interaction between condition, group, and valence). Further, they do not provide enough information to gauge whether the effect of age or disease was *disproportionately* greater in the binding condition[1]. Given that their analysis of each condition separately cannot address key questions in the literature and increases vulnerability to type I errors (Cramer et al., 2016), it would have been better to analyse the data gathered by Borg et al. (2011) as a whole.

Aside from the idiosyncratic nature of the analysis, the paper of Borg et al. (2011) also contains errors in statistical reporting of the kind all-too-common in psychology (Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2015). The F-ratio for the interaction in the picture only condition (referred to in the paper as the 'visual memory task') is reported as follows: "$F[1, 39] = 4.2$, $p = .45$" (pp. 23), and for the binding condition: "$F[2, 39] = 4.14$, $p < .05$" (pp. 23). The first thing to note is the incorrect number of degrees of freedom for the first test, which with a 2×3 mixed design and 42 participants should equal (2, 39). Secondly the reported p-value does not match the reported F-statistic; an F-value of 4.2 with (1, 39) degrees of freedom results in a p-value of approximately[2] 0.047 not the reported .45. With the correct degrees of freedom the p-value is around 0.022. Thus, if the reported F-ratios are

[1] A request for raw data was emailed to the corresponding author of this article on the 5th of January 2016. We have not received a reply to this request.

[2] `http://www.danielsoper.com/statcalc3/calc.aspx?id=7` or the `pf()` function in `R`

to be believed, it appears that the group by valence interaction was significant in *both* the picture only and binding conditions. Those that cite Borg et al. (2011) as evidence for a disproportionate effect of age on binding in WM should be aware of these errors, both in hypothesis testing and statistical reporting.

More recently Fandakova et al. (2014) used the global-local recognition paradigm of Oberauer (2005) to assess WM for letter-location conjunctions in participants of various ages. In this task participants studied six sequentially presented letters in different positions and were then asked to perform a recognition task for the letters presented (global) or the exact pairings of letter and location (local). Structural equation modelling was used to disentangle the contributions of item memory, which is assumed to contribute to both local and global recognition, and item-context memory, assumed to contribute to local recognition only. Fandakova et al. (2014) found that estimates of the item memory latent factor did not significantly differ between their groups of healthy younger and older adults, whereas estimates of the item-context binding factor did ("$\Delta\chi^2 = 56.03$, $df = 1$, $p < .0001$, $d = 1.47$", pp. 145). In addition to this model based analysis, they also report analysis of accuracy in the global and local tasks. The comparison of younger and older adults' performance was significant for both the global ("$t(153.1) = 5.80$, $p < .001$, $d = 0.90$", pp. 144) and local ("$t(168) = 7.74$, $p < .001$, $d = 1.14$", pp. 144) tasks, with a slightly larger effect size for the local comparison. However note that, like Mitchell and colleagues, the vital evidence required for gauging the evidence for a specific binding deficit with healthy ageing in the data of Fandakova et al. (2014) is missing. Granted this was not of primary interest to the authors, who were interested in the relationship between binding efficacy and performance on a broader battery of WM tasks.

There have also been reports that have failed to find a disproportionate effect of healthy ageing on feature-location binding in recognition memory tasks. Bopp and Verhaeghen (2009) present the results of three studies using a paradigm in which participants monitored a stream of stimuli (either letters or a cross in a 4×4 grid) for repetitions. The stimuli were presented in either one or more 'series' which were

differentiated by both their location on the screen and surrounding border colour. In the condition with one such series there is no need to bind the stimuli to their context whereas for more than one series participants had to bind content to context to detect the repetition in each series. In their first experiment they found an interaction between age group and number of series, however, following this up in their second experiment they showed that this was likely due to younger adults' ceiling performance in the single series condition. In an experiment with a greater memory load, thus bringing performance off ceiling, there was no-evidence of the crucial two-way interaction. Thus this study does not support the notion that increasing the task-demand for binding between different contextual features—including location—increases the magnitude of age effects on WM. Nevertheless, whilst location was one of the contextual features used it is still possible that the older group in Bopp and Verhaeghen (2009) were able to use the colour context only, to mask any difficulty binding items to location. Bopp and Verhaeghen (2009) did conduct a third study which removed the location contextual feature leading to poorer performance overall but unfortunately the corresponding experiment removing the colour contextual cue was not performed. It is, therefore, debatable whether the Bopp and Verhaeghen (2009) experiments can be considered as providing evidence bearing on location binding deficits with age.

More recently, however, Read et al. (2016) have provided evidence, using the conventional change detection paradigm, that clearly bears on the question of whether older adults struggle to retain what went where in VWM. Participants studied four coloured squares that were presented either simultaneously or sequentially in different locations and following a 900 ms delay were presented with a single probe item. In one block participants had to indicate whether a change had occurred to either colour or location, whereas in another block they had to indicate whether the conjunction of colour and location depicted in the probe was part of the original set. In this experiment there was no hint of an interaction between age-group and condition ($F < 1$) and no suggestion that the mode of presentation (simultaneous or sequential) modulated this. In a second experiment Read et al. (2016)

went further and introduced shape as an additional feature in order to include trials probing VWM for the binding between surface features as well as surface features to location. The different types of binding change (colour-location, shape-location, colour-shape) were presented within the same block of trials and unfortunately the analysis did not address accuracy differences between these trial types. Nevertheless there was no evidence for an age-related binding deficit, in fact, Read et al. (2016) note that their pattern of means suggested a larger binding cost (relative to the either condition) for younger adults.

Recently the focus has shifted to assessing the recall of item locations from VWM and has suggested that older adults are more likely to re-locate a previously seen feature (e.g. colour) to a location previously occupied by a different feature. That is, older adults are more likely to commit so called 'mis-binding' errors (Peich et al., 2013). Following up on these initial findings, Pertzov et al. (2015) have suggested that, once age differences in the recall of the features themselves are corrected for, older adults do not commit any more mis-binding errors than younger adults. Further research using recall paradigms such as these will be important in giving more fine-grained information regarding the precision of object-location memory across the lifespan, but as of yet do not give any reason to propose specific binding deficits with healthy ageing.

In summary, early work (Mitchell, Johnson, Raye, Mather, & D'Esposito, 2000; Mitchell, Johnson, Raye, & D'Esposito, 2000) along with some more recent follow up studies (Borg et al., 2011; Fandakova et al., 2014) have suggested that healthy adult ageing may bring about changes to one's ability to retain information on what went where in VWM. Thorough critical analysis, however, suggests that this evidence is not as strong as it first appears; evidence for critical interactions is either not sufficient by conventional standards ($p > 0.05$) or is not presented at all. Of course this alone cannot be used to argue *against* age-related location binding deficits, but this question would clearly benefit from more data. Importantly, several studies have been published using both recognition and recall paradigms that—despite having sample sizes approximately similar to, or larger than, previous studies—lend no

support to a specific location binding deficit with age (Bopp & Verhaeghen, 2009; Pertzov et al., 2015; Read et al., 2016).

Here we encounter the problem that failure to reject the null, in the absence of proper power analysis, does not constitute evidence *against* the key interaction effect. Luckily there is enough information in some of the papers assessing age differences in recognition of item-location conjunctions to calculate default Bayes factors for the crucial age × condition interaction. Both Mitchell, Johnson, Raye, Mather, and D'Esposito (2000) and Read et al. (2016) report F ratios with 1 degree of freedom in the numerator (a 2×2 interaction) which can be converted to a t value via the relationship: $t = \sqrt{F}$. For the Mitchell study the condition factor represents the average of object and location only conditions versus the binding condition, whereas for Read et al. this is a comparison of the either and binding conditions. The t values and group sizes reported for each experiment can be converted to the default Bayes factors outlined by Rouder, Speckman, Sun, Morey, and Iverson (2009) using the `ttest.tstat()` function from the `BayesFactor` package (R. D. Morey & Rouder, 2015), thus giving an indication of the strength of evidence for or against the interaction in each experiment. For the two experiments reported by Mitchell and colleagues the evidence is far from compelling; the interaction F value from Experiment 1 (3.67) with 24 in each group results in a Bayes factor of 1.25 in favour of the interaction, whereas for Experiment 2 ($F = 2.38$, 16 each group) $B_{10} = 0.82$ yielding approximately equivalent evidence *against* the interaction ($1/0.82 = 1.21$). The data reported by Read and colleagues are slightly more diagnostic; for their first experiment the F value of 0.75 with 31 in each group corresponds to $B_{10} = 0.35$, preferring the null by approximately 2.8-to-1. As noted by Read et al. (2016), for Experiment 2 ($F = 2.92$, 42 younger and 38 older)[3] the pattern of performance points towards a greater binding cost for younger adults, nevertheless, the Bayes factor very slightly prefers the absence of the age by condition interaction,

[3]On page 10, Read et al. (2016) note that 2 younger adults and 1 older adult were excluded from the analysis for at- or below-chance accuracy. However, the degrees of freedom (1, 78) for the F ratio reported on page 11 are consistent with an analysis of the full data set. While errors in reporting like this are concerning (Nuijten et al., 2015), using the reduced group sizes does not greatly change the Bayes factor.

$B_{10} = 0.82$. In summary, the weight of the existing evidence appears to go against a differential effect of age across change detection tasks assessing VWM for features and their locations. However, it is clear that the magnitude of this evidence is far from convincing and clearly more data are needed on this question. This study aimed to add to the body of work assessing item-location binding in healthy ageing.

5.2 Experiment 7 – Mixed versus Blocked Trials with Colour and Location

The present work was motivated by the findings of Cowan et al. (2006) who, unlike previous studies, reported a clear interaction in their first experiment. Older adults were disproportionately poor at detecting changes to colour-location binding when these changes were mixed in the same trial block with changes to colour only; whereas in a follow-up experiment there was no evidence for an age-related binding deficit when these changes were presented in separate blocks. In the experiments reported in the previous Chapter, we assessed whether mixing changes to colour or shape only along with changes to colour-shape conjunctions would make older adults specifically less sensitive to the binding changes. We did not find this to be the case; in fact Bayes factors pointed fairly strongly towards the absence of crucial age × condition interactions. Given that Experiment 1A of Cowan et al. (2006) is possibly the strongest existing evidence for a deficit of colour-location binding with age (see above), the present study aimed to recreate these findings using an improved paradigm.

Participants were presented with a circular array of 3 or 6 coloured circles and following a brief interval a single test item was presented on which participants had to make a recognition judgement. The probe item could be identical to one of the previously studied items in terms of conjunction of colour and location or one of three different kinds of change could have occurred (see Figure 5.1). The probe could be in a previously occupied location but be a brand new colour (colour change), the probe could appear filled in an old colour but in a location that was not

previously occupied (location change), or an old colour could appear in a location that was previously occupied by a different colour (binding change). Crucially these different kinds of change were seen in separate blocks of trials (along with no-change trials) by some participants, whereas for others they were mixed together with no indication of which kind was likely to occur.

There are good reasons to think that this paradigm is an improvement on that used by Cowan et al. (2006). Firstly, as we restricted location to be selected from a set of 8 surrounding an invisible circle, we were able to meaningfully probe VWM for the locations occupied in the memory array. Cowan and colleagues selected locations at random with some restrictions on spacing between items and were thus unable to do this without placing demands on memory for very fine-grained spatial information. Including trials probing memory for location only may be important as it ensures that participants are motivated to pay attention to each component feature equally (as a change is just as likely to occur for location as it is for colour). Secondly, like Cowan et al. (2006) we had participants make a judgement on a single item, however, unlike their study we do not present un-probed items. Further we selected the colours for each memory array from a set of 8 *without* replacement. The presence of duplicated colours and un-probed items in the test array in the study of Cowan et al. (2006) may have acted as an additional cue as to whether a change had occurred and what kind of change could have occurred (see, Cowan et al., 2014). For colour change trials in their study the probe was always unique whereas for binding change trials the probe was always a duplicate. The present study, by avoiding duplicates and presenting a single item in the probe array, avoids these potential confounds.

Method

Participants

Forty-eight younger adults were recruited from the student population of the University of Edinburgh and 49 healthy older adults were recruited from the Psychology research volunteer panel. None of these individuals had participated in the colour-

Table 5.1: Participant characteristics across the mixed and blocked conditions of Experiment 7

	Blocked		Mixed	
	Younger	Older	Younger	Older
N	24	25	24	24
N_{Female}	18	16	17	17
Mean Age (SD)	20.71 (2.53)	70.00 (4.77)	21.12 (1.73)	71.42 (4.67)
Years of Education	16.02 (1.99)	16.42 (2.93)	16.62 (1.58)	16.58 (3.85)
NART Verbal IQ	108.45 (4.06)	119.42 (3.59)	108.88 (5.71)	119.33 (5.29)
MMSE	-	29.32 (0.85)	-	29.50 (0.93)

shape experiment reported in Chapter 4. Recruits were offered £5 in return for participation for the 45 minute session. Table 5.1 presents the demographic information of the participants split between the mixed and blocked conditions. Once again age-groups were well matched for years of education and the healthy older adults received higher estimates of verbal-IQ from the NART. All older adults scored 27 or above on the MMSE.

Stimuli and Apparatus

Our stimuli in this experiment were made by selecting colours from the set of 8 difficult to name colours developed by Brockmole et al. (2008) and using these colours to fill circles placed in any of 8 locations, like Experiment 6. Stimulus locations appeared every 45° around an invisible circle, with a radius of 2.6° of visual angle, centered in the middle of the screen. Each stimulus circle had a radius of approximately 0.5° of visual angle and was separated from other items centre-to-centre by at least 2°. The experiment was programmed using PsychoPy (Peirce, 2007, 2009) and presented over a 18" E96f+SB ViewSonic monitor with a resolution of 1024×768 refresh rate of 100 Hz.

Design and Procedure

Prior to the main change detection task both groups completed the NART (Nelson, 1982) to obtain an estimate of verbal-IQ (see Table 5.1) and a test of colour vision

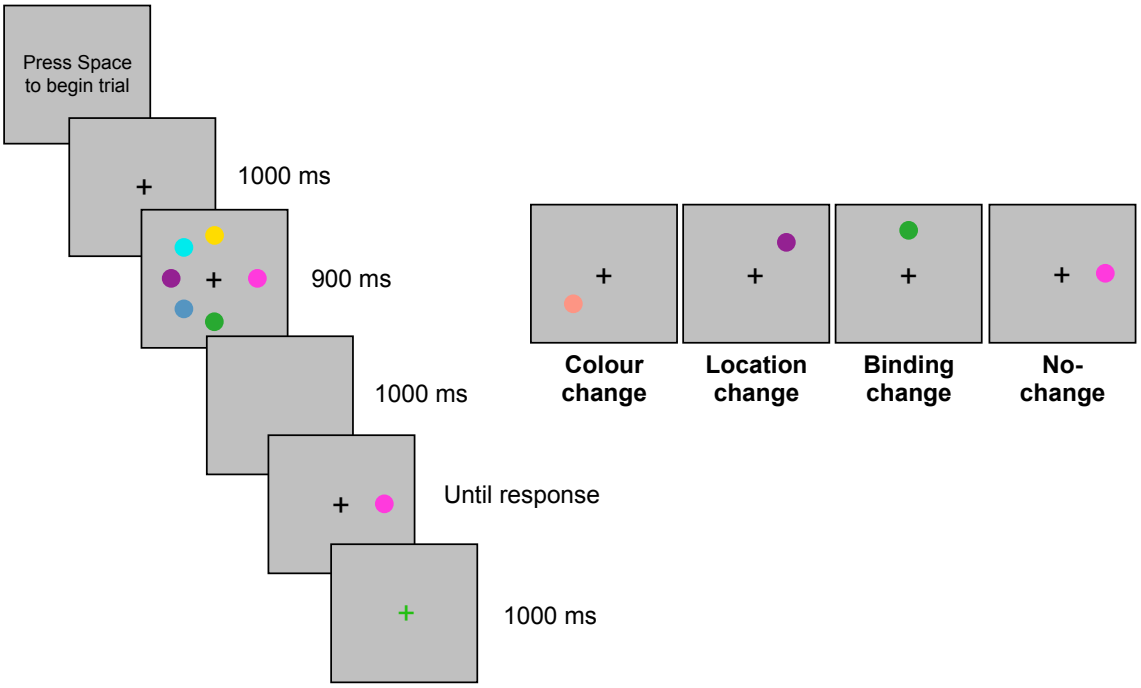

Figure 5.1: Trial sequence and illustrations of different trial types in Experiment 7.

(Dvorine, 1963). The older group also completed the MMSE (Folstein et al., 1975) prior to completing the main part of the experiment.

Figure 5.1 presents the general trial sequence along with examples of the different types of probe arrays encountered in this experiment. Participants initiated each trial by pressing the space-bar and following a 1000 ms fixation cross the memory array appeared for 900 ms. This was followed by a 1000 ms blank retention interval and the central probe item which remained present until a response was made. Finally participants were presented feedback for 1000 ms in the form of a fixation cross that was filled green for a correct response and red otherwise.

Half of the trials presented to participants involved no-change as the probe was selected at random from one of the 3 or 6 objects presented. The remaining half of trials were split between colour change, location change, and binding change types. A colour change involved filling a circle at a previously seen location with a colour from outside the original memory set and a location change involved presenting a circle at a previously unoccupied location filled in an old colour. A binding change involved presenting an old colour at a location that was previously occupied by a different colour. As described above some participants saw these types of change trial

in separate blocks whereas for others they were mixed together with no indication of the kind of trial.

The main experiment was split into 3 blocks with 32 change and 32 no-change trials distributed evenly across the different set sizes. For the blocked condition all change trials were of a single type and for the mixed condition a change was equally likely to occur for colour, location, and binding. Participants in the blocked condition were given 6 practice trials looking for a particular kind of change before the corresponding block whereas participants in the mixed condition were given 18 practice trials before the first block with all three kinds of change trial present. In the blocked condition the order of the three memory conditions (colour, location, binding) was fully counterbalanced.

As in the analysis of Experiment 6, in the present analysis we explore general trends in raw accuracy in the blocked and mixed conditions. This is followed up with a model based analysis of sensitivity and bias.

Results – Raw Accuracy

Blocked Trials

Figure 5.2 presents raw accuracy for each age-group in the blocked condition across the experimental factors of memory condition, set size, and whether or not a change occurred. Visually there do not appear to be any clear differences in the pattern of performance between age-groups in the colour only and binding conditions. In the location condition, however, younger adults show less of an effect of array size—we return to this pattern later. Posterior means (and medians) for each parameter of our logit model are presented in Table 5.2 along with their highest density intervals. As in previous Chapters we use the resulting MCMC chains to construct specific contrasts, allowing us to test hypotheses of interest.

Overall, performance was better in the conditions requiring memory for individual features (colour or location) relative to the binding condition, 0.357 [0.228, 0.485]. Also, unsurprisingly, older adults were less likely to give correct responses relative to younger adults, -0.630 [-0.886, -0.375], and there was a clear effect of in-

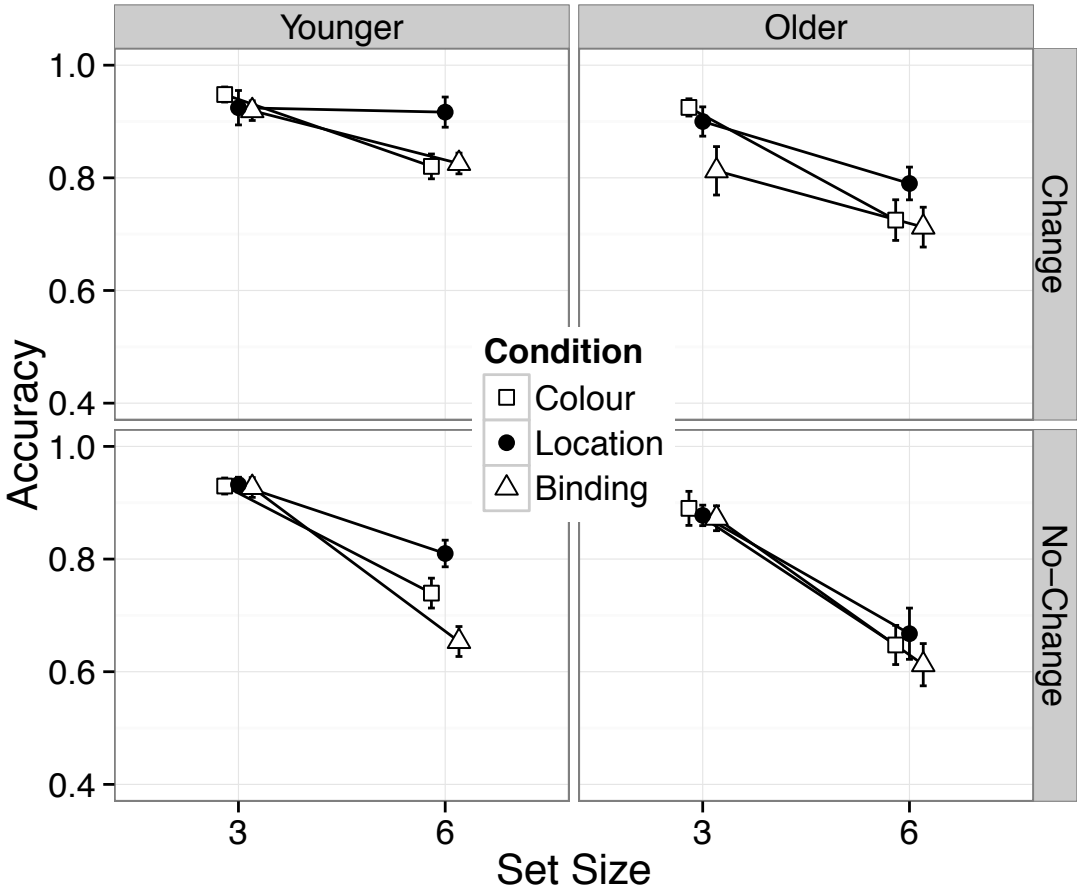

Figure 5.2: Accuracy for blocked trials across age-groups and experimental conditions in Experiment 7. Error bars are ± standard error.

creasing set size, -1.215 [-1.345, -1.092]. Crucially, the contrast between performance in the feature and binding conditions was no larger for older adults than for younger, 0.015 [-0.242, 0.275], with this contrast centered on zero and the HDIs straddling small effect sizes. This interaction was, however, modulated by set size; as shown in Figure 5.2 the discrepancy between feature and binding accuracy is greater for younger adults at set size 6, whereas for older adults it is somewhat larger at set size 3, -0.636 [-1.157, -0.132]. As noted above this is largely driven by the fact that, for younger adults, the effect of increasing the number of to-be-remembered items was less pronounced in the location condition. Thus we repeated these contrasts but focused on the comparison of the colour only and binding conditions (as opposed to the average of the feature conditions versus binding). There was no clear evidence for this three-way interaction when considering only colour and binding,

Table 5.2: Posterior quantities from logit model for the Blocked condition

			95% HDI		
Parameter	Mean	Median	lower	upper	ESS
β_0	1.775	1.774	1.647	1.902	3436.526
$\boldsymbol{\beta}_1$: (1) Location	0.179	0.179	0.089	0.272	12392.361
$\boldsymbol{\beta}_2$: (2) Binding	-0.238	-0.238	-0.323	-0.152	12300.576
$\boldsymbol{\beta}_3$: (3) SS6	-0.608	-0.608	-0.672	-0.546	19626.783
$\boldsymbol{\beta}_4$: (4) Older Group	-0.315	-0.315	-0.443	-0.187	3523.703
$\boldsymbol{\beta}_5$: (5) Change	0.175	0.174	0.112	0.238	19083.220
$\boldsymbol{\beta}_6$: 1×3	0.168	0.168	0.077	0.256	13722.601
$\boldsymbol{\beta}_7$: 2×3	-0.011	-0.011	-0.096	0.074	12558.764
$\boldsymbol{\beta}_8$: 1×4	-0.057	-0.057	-0.150	0.033	12079.410
$\boldsymbol{\beta}_9$: 2×4	-0.005	-0.005	-0.092	0.081	12195.434
$\boldsymbol{\beta}_{10}$: 1×5	0.041	0.041	-0.049	0.132	13563.203
$\boldsymbol{\beta}_{11}$: 2×5	-0.070	-0.069	-0.157	0.015	12683.714
$\boldsymbol{\beta}_{12}$: 3×4	-0.005	-0.006	-0.069	0.056	20317.812
$\boldsymbol{\beta}_{13}$: 3×5	0.150	0.150	0.088	0.214	17520.950
$\boldsymbol{\beta}_{14}$: 4×5	-0.035	-0.035	-0.098	0.028	18471.498
$\boldsymbol{\beta}_{15}$: $1 \times 3 \times 4$	-0.102	-0.102	-0.193	-0.010	13239.213
$\boldsymbol{\beta}_{16}$: $2 \times 3 \times 4$	0.106	0.106	0.022	0.193	12631.022
$\boldsymbol{\beta}_{17}$: $1 \times 3 \times 5$	0.039	0.039	-0.056	0.126	13267.757
$\boldsymbol{\beta}_{18}$: $2 \times 3 \times 5$	0.099	0.098	0.014	0.186	12320.768
$\boldsymbol{\beta}_{19}$: $1 \times 4 \times 5$	0.038	0.038	-0.050	0.131	13393.196
$\boldsymbol{\beta}_{20}$: $2 \times 4 \times 5$	-0.071	-0.071	-0.154	0.018	12338.476
$\boldsymbol{\beta}_{21}$: $3 \times 4 \times 5$	-0.044	-0.044	-0.107	0.019	19844.658
$\boldsymbol{\beta}_{22}$: $1 \times 3 \times 4 \times 5$	-0.043	-0.043	-0.130	0.050	12542.040
$\boldsymbol{\beta}_{23}$: $2 \times 3 \times 4 \times 5$	0.027	0.027	-0.060	0.111	13016.536
σ_s	0.392	0.388	0.294	0.499	10589.907

Note: The effects coded variables were as follows: (1) Location = 1, Binding = 0, Colour = -1, (2) Location = 0, Binding = 1, Colour = -1, (3) SS3 = -1, SS6 = 1, (4) Younger = -1, Older = 1, (5) No-Change = -1, Change = 1. Interaction contrasts were products of these effects coded variables.

-0.439 [-1.041, 0.160].

There is a simple explanation as to why our younger group displayed a smaller set size effect in the location condition. As location was constrained and participants in the blocked condition were aware of what kind of change was possible on a given trial an appropriate strategy was to note the empty locations in the memory array; if, when the probe appeared, the locations were still empty then there had been no change, whereas if one were occupied then a location change must have occurred. Using this strategy would make larger arrays easier as there are fewer empty locations

to monitor. It is interesting to note that, overall, younger adults were more likely to note this aspect of the task as evidenced by their pattern of performance and post-experiment discussion with the researcher. Assessing such strategy difference between age-groups on working memory tasks will prove important in future work to separate out true effects of healthy ageing from difference in strategy use (Logie et al., 2015). Finally, in the previous Chapter assessing VWM for conjunctions of colour and shape there was evidence that the effect of age was larger for trials containing a change relative to trials involving no-change. This was also the case here, as the age $\times$ trial type contrast was of a similar magnitude the previous experiment, suggesting a smaller age effect for no-change trials, -0.352 [-0.478, -0.230].

In summary, the pattern of results depicted in Figure 5.2 and the parameter estimates presented in Table 5.2 give no reason to believe that older adults specifically struggled to perform our change detection task when it required them to retain conjunctions of colour and location. However, given the findings of Cowan et al. (2006), we may expect mixing together trials containing changes to individual features and changes to conjunctions to reveal a specific deficit.

Mixed Trials

The accuracy of participants in the mixed condition is presented in Figure 5.3. Visually it appears that accuracy on no-change trials is not greatly affected by age, whereas older adults are less accurate on change trials relative to younger adults. Further, in both groups accuracy in detecting binding changes appears to be more-or-less the same as accuracy for changes to colour only, which would not be expected if our older adults were experiencing specific difficulty in detecting alterations to the pairing of colour and location.

Table 5.3 presents a summary of parameter estimates from the analysis of the mixed condition. Contrasts revealed that, like the blocked condition, accuracy was better for features relative to binding, 0.403 [0.216, 0.592], older adults were less likely to produce correct responses, -0.707 [-1.015, -0.394], and increasing the array size degraded performance, -1.085 [-1.236, -0.940]. Contrasting feature and bind-

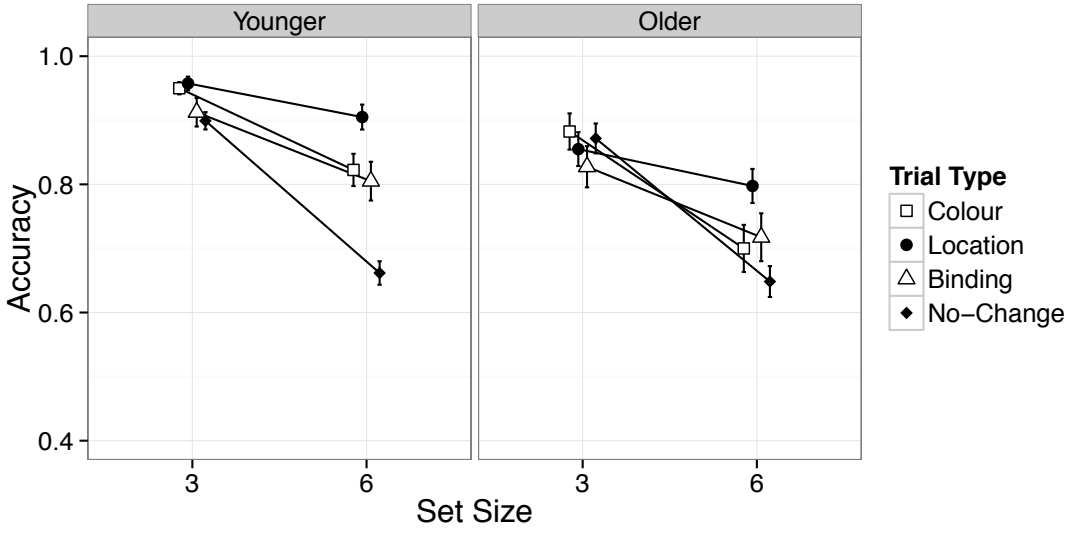

Figure 5.3: Accuracy for mixed trials across age-groups and experimental conditions in Experiment 7. Error bars are ± standard error.

ing change detection revealed that, if anything, the discrepancy was larger in the younger group, -0.225 [-0.598, 0.158], although the HDI limits clearly overlap zero. Restricting this contrast to the difference between the colour and binding conditions bring the mean closer to zero, -0.077 [-0.513, 0.368], and still favours a smaller cost in the younger group. There was no suggestion that set size modulated this either when comparing features against binding, 0.037 [-0.724, 0.786], or colour only against binding, -0.138 [-1.012, 0.752]. Comparing accuracy on trials where there was a change (an average of colour, location, and binding parameters) to accuracy when there was no-change we find that a correct response was more likely for the former than the latter, 0.531 [0.410, 0.656]. This was qualified, however, by a clear interaction with age-group such that the difference between change and no-change detection was far less pronounced in the older group, -0.352 [-0.478, -0.230]. As shown in Figure 5.3 older adults are less likely to detect a change compared to younger adults, whereas for detecting sameness the age-effect is negligible.

In summary, the mixed data—like accuracy in the blocked condition—do not give any indication that healthy older adults struggle to retain the correct binding between colour and location. The next analysis follows this up with a direct contrast of change detection across the two block types.

Table 5.3: Posterior quantities from logit model for the Mixed condition

			95% HDI		
Parameter	Mean	Median	lower	upper	ESS
β_0	1.789	1.789	1.636	1.947	2295.380
$\boldsymbol{\beta}_1$: (1) Location	0.411	0.410	0.264	0.556	9169.901
$\boldsymbol{\beta}_2$: (2) Binding	-0.136	-0.137	-0.258	-0.009	12920.896
$\boldsymbol{\beta}_3$: (3) No-Change	-0.398	-0.398	-0.492	-0.308	9531.336
$\boldsymbol{\beta}_4$: (4) SS6	-0.543	-0.542	-0.618	-0.470	9191.074
$\boldsymbol{\beta}_5$: (5) Older Group	-0.354	-0.354	-0.507	-0.197	2153.631
$\boldsymbol{\beta}_6$: 1×4	0.196	0.196	0.048	0.341	8943.044
$\boldsymbol{\beta}_7$: 2×4	0.132	0.133	0.007	0.259	13329.995
$\boldsymbol{\beta}_8$: 3×4	-0.184	-0.184	-0.280	-0.095	8505.681
$\boldsymbol{\beta}_9$: 1×5	-0.200	-0.198	-0.345	-0.051	9114.740
$\boldsymbol{\beta}_{10}$: 2×5	-0.013	-0.013	-0.137	0.113	13122.054
$\boldsymbol{\beta}_{11}$: 3×5	0.264	0.264	0.173	0.359	9322.522
$\boldsymbol{\beta}_{12}$: 4×5	0.073	0.073	0.000	0.147	8898.226
$\boldsymbol{\beta}_{13}$: $1 \times 4 \times 5$	0.056	0.056	-0.092	0.203	8919.237
$\boldsymbol{\beta}_{14}$: $2 \times 4 \times 5$	0.002	0.002	-0.121	0.129	13160.935
$\boldsymbol{\beta}_{15}$: $3 \times 4 \times 5$	-0.026	-0.026	-0.117	0.067	9098.443
σ_s	0.473	0.468	0.360	0.594	13109.712

Note: The effects coded variables were as follows: (1) Location = 1, Binding = 0, No-Change = 0, Colour = -1, (2) Location = 0, Binding = 1, No-Change = 0, Colour = -1, (3) Location = 0, Binding = 0, No-Change = 1, Colour = -1, (4) SS3 = -1, SS6 = 1, (5) Younger = -1, Older = 1. Interaction contrasts were products of these effects coded variables.

Mixed Versus Blocked

The individual analyses presented above give no reason to suspect that older adults experience difficulty in binding different colours to the exact locations they were presented in. However, it is insufficient to assess potentially important differences in our participants' ability to detect changes when different change types were mixed as opposed to when they were separated. Accuracy on change trials was combined and the Bayesian logistic ANOVA (see Chapter 2) was estimated with the factors of change type (colour, location, binding), set size (3, 6), age-group (young, old), and block-type (mixed, blocked). A summary of the resulting parameter estimates is given in Table 5.4.

The key findings from the separate analyses reported above were recreated in this analysis of the large, combined data set. Change detection performance was much

Table 5.4: Posterior quantities from logit model comparing change trials in the mixed and blocked conditions

			95% HDI		
Parameter	Mean	Median	lower	upper	ESS
β_0	2.018	2.018	1.866	2.167	2729.411
$\boldsymbol{\beta}_1$: (1) Location	0.259	0.259	0.162	0.360	12559.754
$\boldsymbol{\beta}_2$: (2) Binding	-0.296	-0.297	-0.387	-0.208	13514.816
$\boldsymbol{\beta}_3$: (3) SS6	-0.484	-0.484	-0.552	-0.418	19886.761
$\boldsymbol{\beta}_4$: (4) Older Group	-0.408	-0.408	-0.562	-0.254	2736.175
$\boldsymbol{\beta}_5$: (5) Mixed	-0.006	-0.006	-0.157	0.147	2667.493
$\boldsymbol{\beta}_6$: 1×3	0.177	0.177	0.079	0.275	12019.178
$\boldsymbol{\beta}_7$: 2×3	0.080	0.080	-0.008	0.170	13270.580
$\boldsymbol{\beta}_8$: 1×4	-0.066	-0.066	-0.165	0.031	12959.336
$\boldsymbol{\beta}_9$: 2×4	-0.002	-0.002	-0.094	0.085	13387.035
$\boldsymbol{\beta}_{10}$: 1×5	0.030	0.030	-0.067	0.129	11666.722
$\boldsymbol{\beta}_{11}$: 2×5	0.020	0.020	-0.068	0.112	12580.812
$\boldsymbol{\beta}_{12}$: 3×4	0.014	0.014	-0.055	0.079	19355.098
$\boldsymbol{\beta}_{13}$: 3×5	-0.014	-0.014	-0.080	0.055	19059.012
$\boldsymbol{\beta}_{14}$: 4×5	-0.059	-0.059	-0.211	0.095	2566.470
$\boldsymbol{\beta}_{15}$: $1 \times 3 \times 4$	-0.050	-0.050	-0.147	0.048	13285.550
$\boldsymbol{\beta}_{16}$: $2 \times 3 \times 4$	0.066	0.066	-0.024	0.153	13288.505
$\boldsymbol{\beta}_{17}$: $1 \times 3 \times 5$	-0.036	-0.035	-0.133	0.062	13456.869
$\boldsymbol{\beta}_{18}$: $2 \times 3 \times 5$	-0.008	-0.008	-0.097	0.081	13824.915
$\boldsymbol{\beta}_{19}$: $1 \times 4 \times 5$	-0.049	-0.049	-0.147	0.051	11452.280
$\boldsymbol{\beta}_{20}$: $2 \times 4 \times 5$	0.079	0.079	-0.011	0.169	12187.668
$\boldsymbol{\beta}_{21}$: $3 \times 4 \times 5$	0.068	0.068	-0.001	0.134	19925.194
$\boldsymbol{\beta}_{22}$: $1 \times 3 \times 4 \times 5$	0.098	0.098	-0.001	0.197	12323.468
$\boldsymbol{\beta}_{23}$: $2 \times 3 \times 4 \times 5$	-0.071	-0.071	-0.160	0.020	12798.248
σ_s	0.669	0.666	0.554	0.795	11302.557

Note: The effects coded variables were as follows: (1) Location = 1, Binding = 0, Colour = -1, (2) Location = 0, Binding = 1, Colour = -1, (3) SS3 = -1, SS6 = 1, (4) Younger = -1, Older = 1, (5) Blocked = -1, Mixed = 1. Interaction contrasts were products of these effects coded variables.

better when a change occurred to an individual feature relative to the conjunction of features, 0.445 [0.312, 0.581], older adults were less likely to detect changes overall, -0.816 [-1.123, -0.508], and set size had a very large effect -0.968 [-1.104, -0.837]. Further, the difference between accuracy for colour and binding changes was not mediated by age-group, 0.006 [-0.256, 0.282].

This combined analysis revealed that, overall, performance did not differ depending on whether different trial types were mixed or blocked, -0.013 [-0.314, 0.294]. The

crucial question is whether or not the manner in which change trials were presented affected older adults' binding change detection specifically. Thus we constructed a specific contrast comparing the size of the binding cost (average of the feature conditions versus binding) across age-groups in the mixed and blocked conditions; this revealed that there was no clear role of block type, -0.471 [-1.011, 0.063]. The direction of this contrast suggests that, if anything, age-differences in the binding cost were actually larger in the blocked condition than in the mixed condition. This can also be seen by comparing the top two panels of Figure 5.2 to the change trials in Figure 5.3. Finally, there was no strong evidence for a four-way interaction including set size, 0.637 [-0.177, 1.444]. Given the width of the HDIs the precision afforded to us by the data is not enough to make a clear conclusion regarding this interaction. However, it is important to note that the appearance of such an interaction would not be expected on the basis of Cowan et al's findings and would be difficult to explain in terms of a specific binding deficit with healthy ageing.

Sensitivity and Bias

Analysis of raw accuracy gave no suggestion of a disproporionate effect of age for binding change detection. Further, mixing different varieties of change trial did not clearly modulate this. In the following analysis we probe deeper into these patterns of accuracy by separating the contribution of sensitivity and bias to change detection responses.

Sensitivity

Estimates of sensitivity (P_r) for each age-group in the blocked and mixed conditions across the other experimental factors are presented in Figure 5.4. All four panels show, in essence, the same pattern of performance with the exception that overall sensitivity is lower in the older group (right panels) and younger adults' sensitivity to location changes was high at set size 6 in the blocked condition (top left panel). Tables 5.5 and 5.6 present the results of Bayesian ANOVAs on sensitivity for the blocked and mixed conditions, respectively.

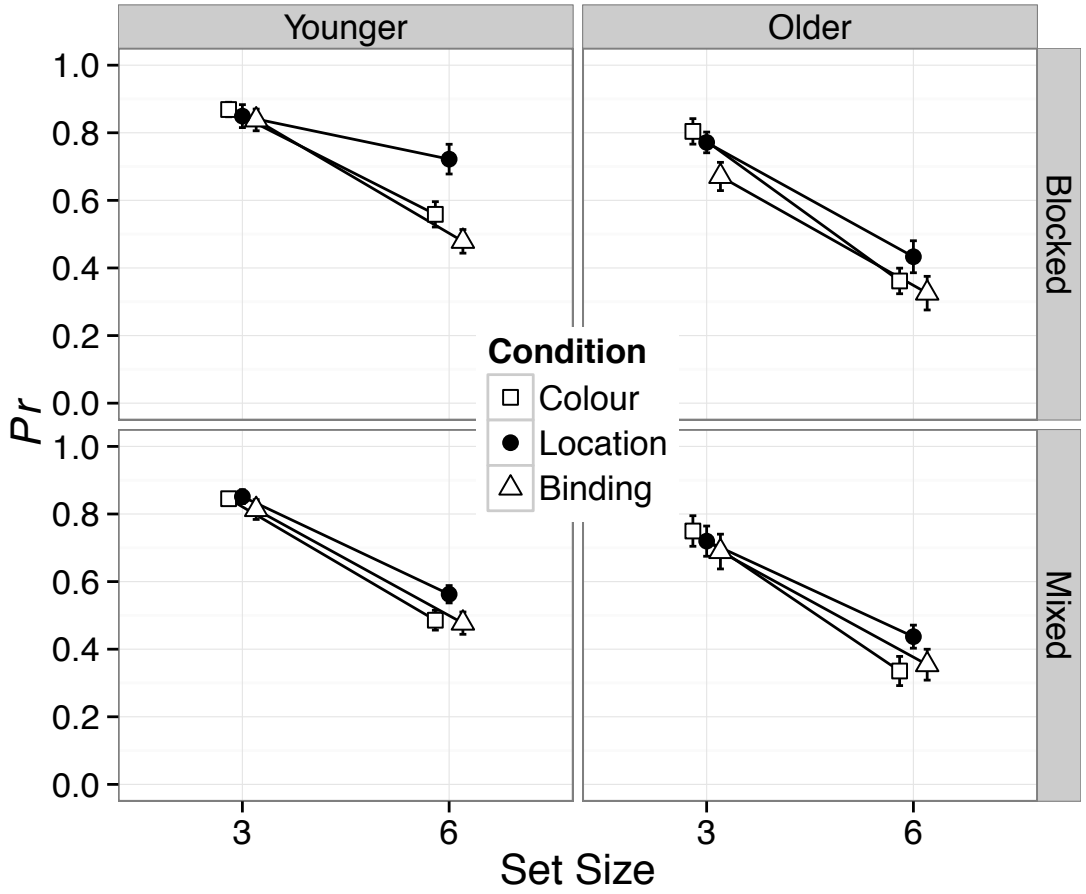

Figure 5.4: P_r (corrected recognition) across age groups and experimental conditions in Experiment 7. Error bars are ± standard error.

In our Bayesian analysis of the blocked condition the winning model included all main effects (memory condition, set size, age-group) as well as condition × set size and age-group × set size interactions (see Table 5.5). Model 5 was identical except for also including the two-way interaction between age-group and memory condition and comparing models 1 and 5 revealed good evidence against this critical interaction ($B_{1,5} = 8.73$). In contrast, there was good evidence *for* the interaction between age and set size ($B_{1,3} = 6.39$), with a larger effect of increasing the number of to-be-remembered objects in the older group.

The winning model from the BANOVA on the mixed data also did not include the age × condition interaction. As Table 5.6 shows, this model included all main effects and the interaction between memory condition and set size. Again, in this analysis model 5 was identical to the first model with the addition of our interaction

Table 5.5: Log Bayes factors for analysis of sensitivity (P_r) in the blocked condition

	Model	$\log(B_{M,0})$	% error
1	$P_r \sim$ C + SS + C:SS + AG + SS:AG + ID	95.33	1.31
2	$P_r \sim$ C + SS + AG + SS:AG + ID	93.66	1.17
3	$P_r \sim$ C + SS + C:SS + AG + ID	93.48	0.78
4	$P_r \sim$ C + SS + C:SS + AG + C:AG + SS:AG + C:SS:AG + ID	93.22	1.02
5	$P_r \sim$ C + SS + C:SS + AG + C:AG + SS:AG + ID	93.16	1.53
6	$P_r \sim$ C + SS + AG + ID	91.93	0.51
7	$P_r \sim$ C + SS + AG + C:AG + SS:AG + ID	91.47	1.24
8	$P_r \sim$ C + SS + C:SS + AG + C:AG + ID	91.27	0.74
9	$P_r \sim$ C + SS + AG + C:AG + ID	89.71	0.62
10	$P_r \sim$ SS + AG + SS:AG + ID	86.82	0.86
11	$P_r \sim$ C + SS + C:SS + ID	86.18	0.60
12	$P_r \sim$ SS + AG + ID	85.34	1.26
13	$P_r \sim$ C + SS + ID	84.66	0.53
14	$P_r \sim$ SS + ID	78.19	0.31
15	$P_r \sim$ C + AG + ID	7.24	0.92
16	$P_r \sim$ AG + ID	5.39	0.45
17	$P_r \sim$ C + AG + C:AG + ID	4.88	4.20
18	$P_r \sim$ C + ID	1.70	0.91

Note: All Bayes factors are relative to a null model with a random participant effect only (model 0: $P_r \sim$ ID). AG = Age-Group, C = Condition, SS = Set Size, ID = participant ID, and ':' denotes an interaction effect

of interest. Comparing these two revealed that the absence of the age × condition interaction was favoured by approximately 14-to-1. This suggests that, if anything, the evidence against a differential effect of age across our memory condition was stronger in the mixed condition than in the blocked condition.

We followed this up with an analysis of the whole data set with an additional factor of block type. This analysis compared a full model to reduced models omitting a single component (main- or interaction-effect) at a time, thus reducing the number of models to be computed (see Brown et al., 2016, for the same approach).

This full analysis revealed that omitting the interaction between memory condition and age-group resulted in a model that, given the data, was over 17 times more likely than the full model. This constitutes strong evidence *against* the suggestion of a specific age-related deficit in binding colour to location in VWM. Further, there was strong evidence against the equally crucial three-way interaction between age × memory condition × block type ($B_{R,F} = 11.397$). For the four-way interaction the evidence was far from compelling ($B_{R,F} = 1.563$); this was likely due to younger adults better detection of location changes in the blocked condition, specifically for set size 6 (see above). When location is omitted the evidence against the four-way

Table 5.6: Log Bayes factors for analysis of sensitivity (P_r) in the mixed condition

	Model	$\log(B_{M,0})$	% error
1	$P_r \sim$ C + SS + C:SS + AG + ID	160.02	1.12
2	$P_r \sim$ C + SS + C:SS + AG + SS:AG + ID	158.44	1.05
3	$P_r \sim$ C + SS + AG + ID	158.15	1.34
4	$P_r \sim$ C + SS + C:SS + ID	158.01	1.37
5	$P_r \sim$ C + SS + C:SS + AG + C:AG + ID	157.35	1.19
6	$P_r \sim$ C + SS + AG + SS:AG + ID	156.61	1.18
7	$P_r \sim$ C + SS + ID	156.14	0.43
8	$P_r \sim$ C + SS + C:SS + AG + C:AG + SS:AG + ID	155.81	2.42
9	$P_r \sim$ C + SS + AG + C:AG + ID	155.49	1.00
10	$P_r \sim$ SS + AG + ID	154.93	0.84
11	$P_r \sim$ C + SS + C:SS + AG + C:AG + SS:AG + C:SS:AG + ID	154.17	3.82
12	$P_r \sim$ C + SS + AG + C:AG + SS:AG + ID	153.94	1.60
13	$P_r \sim$ SS + AG + SS:AG + ID	153.35	2.25
14	$P_r \sim$ SS + ID	152.88	0.47
15	$P_r \sim$ AG + ID	1.82	0.34
16	$P_r \sim$ C + AG + ID	0.26	0.45
17	$P_r \sim$ C + ID	-1.56	0.50
18	$P_r \sim$ C + AG + C:AG + ID	-2.42	0.60

Note: All Bayes factors are relative to a null model with a random participant effect only (model 0: $P_r \sim$ ID). AG = Age-Group, C = Condition, SS = Set Size, ID = participant ID, and ':' denotes an interaction effect

interaction becomes slightly more convincing ($B_{R,F} = 2.699$).

There was overwhelming evidence for an effect of age on change detection sensitivity, ($B_{F,R} = 1.9087 \times 10^4$), but the weight of evidence was against modulation of this age effect by block type, ($B_{R,F} = 3.138$). Thus, regardless of whether participants are required to retain features or feature bindings, mixing different types of trial together within the same task does not disproportionately affect older adults' sensitivity to changes, as was found in Experiment 6. Finally, there was support for an interaction between age and set size, ($B_{F,R} = 4$). As can be seen in Figure 5.4 the effect of group on sensitivity was larger when six to-be-remembered items were presented.

Bias

Response bias in the mixed and blocked conditions is presented in Figure 5.5. Separate Bayesian ANOVAs were conducted on B_r in these two conditions. For the blocked condition the 'winning' model included a main effect of set size *only* (see Table 5.7). In order to gauge the weight of evidence against age differences in response bias we compared the winning model to model 3 which included the main

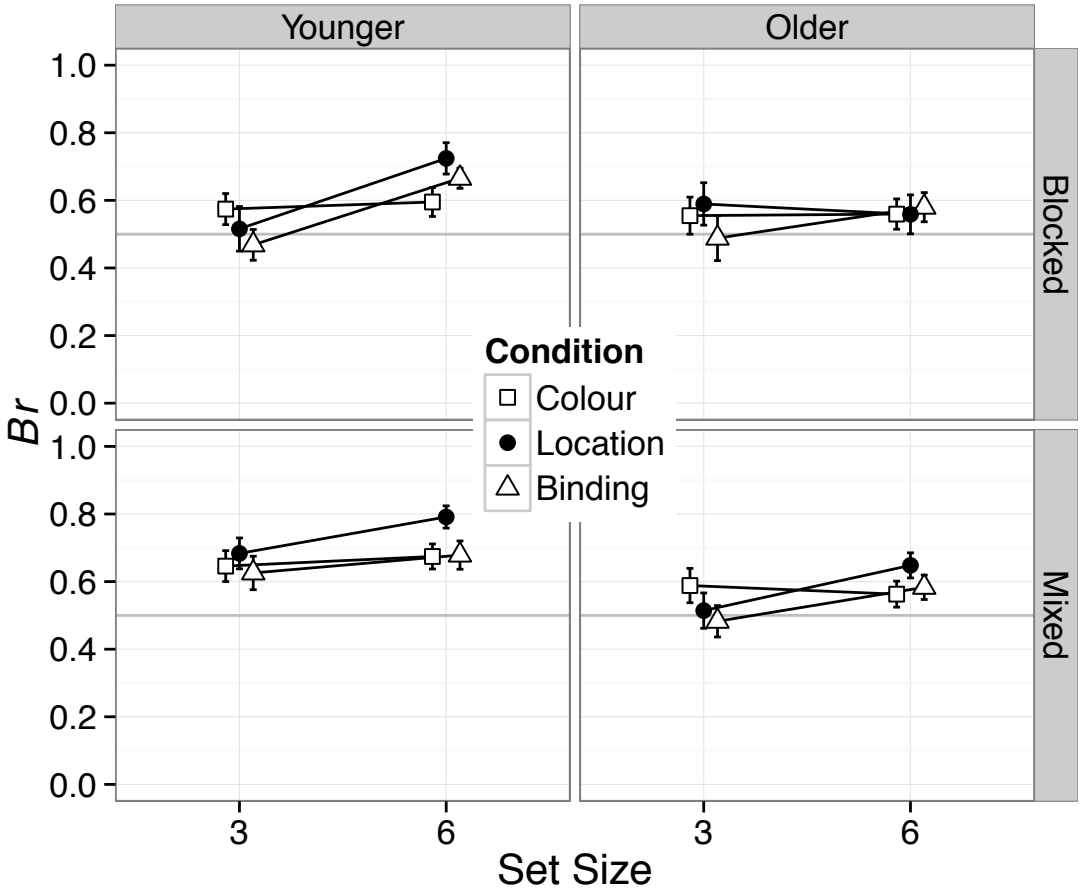

Figure 5.5: B_r across age groups and experimental conditions in Experiment 7. Error bars are ± standard error.

effect of age-group, this revealed that the absence of this effect was favoured by a factor of over 3 ($B_{1,3} = 3.697$). While not entirely necessary, given the absence of an overall effect of age, we also assessed the evidence against an age × condition interaction in the blocked condition. As can be seen in Figure 5.5 there is clearly no such interaction in this data ($B_{8,14} = 14.233$).

The full results of the analysis of the mixed condition are given in Table 5.7. In this case the winning model included an effect of age-group in addition to the effect of set size. As can be seen in Figure 5.5 there was a clear effect of age on response bias ($B_{1,8} = 9.201$) such that younger adults exhibited a more liberal guessing bias whereas older adults were relatively neutral. Model 2 contained all three main effects and model 6 contained, in addition, the crucial age × condition interaction; comparing these models yielded over 6-to-1 evidence against this interaction ($B_{2,6}$

Table 5.7: Log Bayes factors for analysis of guessing bias (B_r) in the blocked condition

	Model	$\log(B_{M,0})$	% error
1	$B_r \sim$ SS + ID	2.26	0.26
2	$B_r \sim$ SS + AG + SS:AG + ID	1.56	4.52
3	$B_r \sim$ SS + AG + ID	0.95	0.68
4	$B_r \sim$ C + SS + ID	-0.14	0.93
5	$B_r \sim$ C + SS + AG + SS:AG + ID	-0.82	2.90
6	$B_r \sim$ C + SS + C:SS + ID	-1.11	1.63
7	$B_r \sim$ AG + ID	-1.32	0.52
8	$B_r \sim$ C + SS + AG + ID	-1.44	1.04
9	$B_r \sim$ C + SS + C:SS + AG + SS:AG + ID	-1.85	1.65
10	$B_r \sim$ C + SS + C:SS + AG + ID	-2.42	1.11
11	$B_r \sim$ C + ID	-2.42	0.26
12	$B_r \sim$ C + SS + AG + C:AG + SS:AG + ID	-3.51	0.79
13	$B_r \sim$ C + AG + ID	-3.75	0.34
14	$B_r \sim$ C + SS + AG + C:AG + ID	-4.10	0.81
15	$B_r \sim$ C + SS + C:SS + AG + C:AG + SS:AG + ID	-4.47	1.30
16	$B_r \sim$ C + SS + C:SS + AG + C:AG + ID	-5.09	0.91
17	$B_r \sim$ C + SS + C:SS + AG + C:AG + SS:AG + C:SS:AG + ID	-5.48	1.41
18	$B_r \sim$ C + AG + C:AG + ID	-6.40	0.51

Note: All Bayes factors are relative to a null model with a random participant effect only (model 0: $B_r \sim$ ID). AG = Age-Group, C = Condition, SS = Set Size, ID = participant ID, and ':' denotes an interaction effect

= 6.391).

The analysis of the full data set proceeded in a slightly different fashion by comparing reduced models to a full, saturated model. This showed substantial evidence for an overall effect of age on response bias ($B_{F,R} = 4.099$) with younger adults exhibiting a greater bias towards responding 'change'. Figure 5.5 appears to show a larger effect of age on response bias in the mixed condition; however the data did not clearly support (or refute) the age-group × block type interaction ($B_{R,F} = 1.988$). Most importantly there was strong evidence against the age × condition ($B_{R,F} = 16.662$) and age × condition × block type ($B_{R,F} = 14.012$) interactions. Finally, while not necessarily crucial for testing suggestions of an age related binding deficit, the omission of the four-way interaction was favoured over its inclusion in the full model ($B_{R,F} = 2.49$).

In summary, younger adults exhibited a tendency towards guessing change when in an uncertain state whereas older adults appeared to be more neutral. The Bayes factors presented above suggest that this was not modulated by the requirement to remember features or their combination nor was there a clear effect of mixing

Table 5.8: Log Bayes factors for analysis of guessing bias (B_r) in the mixed condition

	Model	$\log(B_{M,0})$	% error
1	$B_r \sim$ SS + AG + ID	4.81	0.37
2	$B_r \sim$ C + SS + AG + ID	4.65	1.89
3	$B_r \sim$ C + SS + C:SS + AG + ID	4.37	0.94
4	$B_r \sim$ SS + AG + SS:AG + ID	3.08	1.05
5	$B_r \sim$ C + SS + AG + SS:AG + ID	2.94	0.96
6	$B_r \sim$ C + SS + AG + C:AG + ID	2.79	0.64
7	$B_r \sim$ C + SS + C:SS + AG + SS:AG + ID	2.72	3.52
8	$B_r \sim$ SS + ID	2.60	0.31
9	$B_r \sim$ C + SS + C:SS + AG + C:AG + ID	2.55	1.65
10	$B_r \sim$ C + SS + ID	2.44	0.41
11	$B_r \sim$ AG + ID	2.18	0.73
12	$B_r \sim$ C + SS + C:SS + ID	2.17	1.21
13	$B_r \sim$ C + AG + ID	1.92	0.42
14	$B_r \sim$ C + SS + AG + C:AG + SS:AG + ID	1.08	0.82
15	$B_r \sim$ C + SS + C:SS + AG + C:AG + SS:AG + ID	0.81	0.82
16	$B_r \sim$ C + AG + C:AG + ID	0.04	0.56
17	$B_r \sim$ C + ID	-0.27	0.26
18	$B_r \sim$ C + SS + C:SS + AG + C:AG + SS:AG + C:SS:AG + ID	-0.96	2.82

Note: All Bayes factors are relative to a null model with a random participant effect only (model 0: $B_r \sim$ ID). AG = Age-Group, C = Condition, SS = Set Size, ID = participant ID, and ':' denotes an interaction effect

different trial types.

Discussion

The present results, in terms of our analyses of raw accuracy along with measures of sensitivity and bias, strongly argue against a *specific* age-related difficulty in retaining colour-location conjunctions in VWM. This is perhaps surprising given the repeated assertions that maintaining *what went where* in WM may exhibit disproportionate decline with age (e.g., R. J. Allen, Brown, & Niven, 2013; Brockmole et al., 2008; Borg et al., 2011; Mitchell, Johnson, Raye, Mather, & D'Esposito, 2000). We discuss potential reasons for this discrepancy in the General Discussion.

Specifically the present findings contrast most starkly with those of Cowan et al. (2006) who observed a large age-related binding deficit when different types of trial were mixed and no clear deficit when they were blocked. Here we did not observe this pattern of results as under both mixed and blocked conditions older adults exhibited generally poorer change detection accuracy that was roughly equal across all memory conditions, regardless of whether VWM for features or conjunctions was required (see Figures 5.2, 5.3, and 5.4). If anything the evidence against the

age by condition interaction was clearest in the mixed condition. As noted in the Introduction there are a number of methodological differences between the present study and that of Cowan et al. (2006) made in light of increased understanding of the change detection task and ways of probing VWM (Cowan et al., 2014). One major difference was the inclusion of trials probing recognition of the locations occupied in the memory array, regardless of the corresponding colours presented at these locations. It is possible that in the study of Cowan et al. (2006) participants were biased towards focusing on the colours presented given that 50% of the changes that would occur in a mixed block of trials would be to colour. Allocating greater attention to colour may come at the expense of precise representation of location information consequently reducing discrimination of binding changes (Woodman & Vogel, 2008). If older adults were more likely to engage in this strategic trade off this could conceivably result in the poorer sensitivity to binding changes observed by Cowan et al. in their mixed condition.

Of course this relies on a number of tentative assumptions of which one or more may be untrue. Therefore, we decided to replicate the mixed condition of our study while omitting trials on which location only changed. If we find evidence of a specific age-related binding deficit here then an explanation of Cowan et al. (2006)'s findings may be a strategic trade off induced by including only one kind of feature change. Otherwise, if such a deficit is not found, this serves to replicate our present findings and rule out a potential source of the discrepancy between the two studies.

5.3 Experiment 8 – Omitting Location Only Changes

Experiment 8 aimed to recreate the findings of Experiment 7 with the omission of location change trials, bringing the design closer to that of Cowan et al. (2006).

Method

Participants

Twenty-four younger adults (mean age = 20.96, SD = 3.10) and 24 older adults (71.13, 4.13) were recruited from the same populations as Experiment 7. Participants were offered £5 for completion of the 45 minute session. Thirteen of participants in the older group had taken part in Experiment 6 assessing colour-shape binding reported in Chapter 4 approximately a year previously.

Stimuli, Apparatus, and Procedure

The methodology of Experiment 8 was identical to that of the mixed condition of Experiment 7 with the exception that trials on which location changed were omitted. Participants were informed that changes could occur to colour only (probe could be brand new) or to the combination of colour and location (old colour presented at a location previously occupied by a different colour) and were presented with examples of such trials. As the location trials were not included in this study the number of trial blocks was reduced to 2 each containing 64 trials following 12 practice trials. After completing this experiment participants went on to complete an additional task, the results of which are presented in Chapter 6.

Results

Accuracy across the three kinds of trial—colour change, binding change, or no-change—and two set sizes is presented in Figure 5.6 for each age group. The logistic ANOVA model was estimated with effects coded variables representing main effects of age-group, set-size, and trial type as well as their interactions (see Table 5.9).

Specific contrasts were constructed using the MCMC chains resulting from our model estimation and the results were similar to the analysis of the mixed condition of Experiment 7. There was a clear difference in accuracy for colour versus binding changes, 0.584 [0.329, 0.845], and for set size 6 relative to set size 3, -1.353 [-1.536, -1.166]. As repeatedly shown, older adults were far less likely to produce a correct

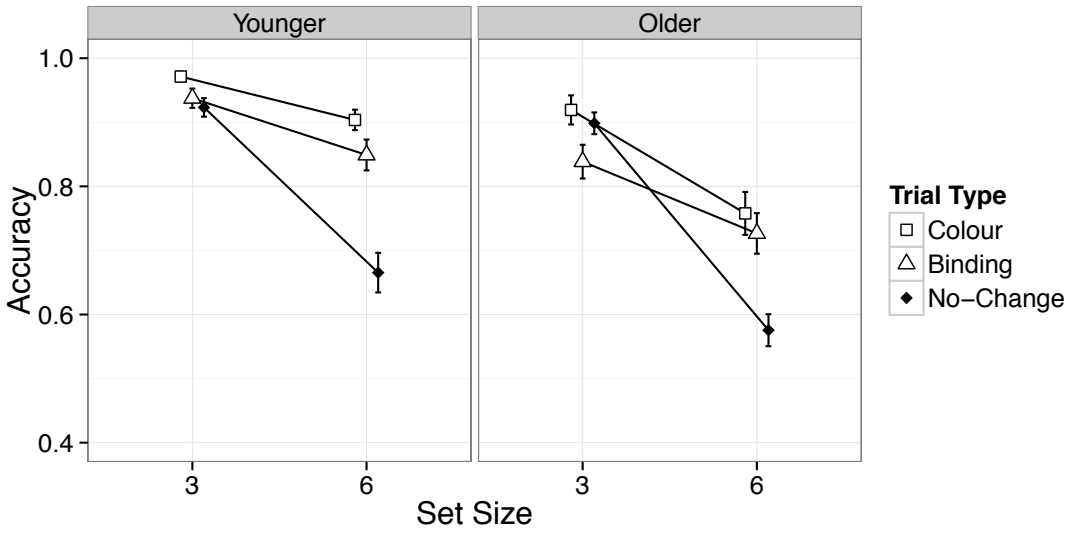

Figure 5.6: Accuracy in Experiment 8 across age-groups and experimental conditions. Error bars are ± standard error.

response on this task than younger adults, -0.822 [-1.140, -0.489]. Crucially, however, the contrast between colour and binding accuracy was not convincingly larger for the older adults than it was for the younger group, -0.099 [-0.354, 0.161]. The HDIs for this contrast span a range of fairly small effect sizes, lending support to the idea that any interaction between age and condition must be rather small (see also, Brown et al., 2016). Finally, there was no indication that set size modulated the two-way interaction between age and condition, -0.320 [-1.352, 0.709], and as shown in Figure 5.6 the difference between colour and binding accuracy for our older adults is slightly larger at set size 3. On the basis of raw accuracy, then, there appears to be no indication that removing trials on which only location changed affected performance. To follow this up separate analysis were conducted on sensitivity and bias.

Sensitivity

Mean estimates of sensitivity (P_r) for Experiment 8 are presented in Figure 5.7. At first glance there appears to be no differential effect of age on the two conditions but a larger effect of set size in the older group. A Bayesian ANOVA tested this initial impression (see Table 5.10). As can be seen in the output of this analysis the

Table 5.9: Posterior quantities from logit model for Experiment 8

			95% HDI		
Parameter	Mean	Median	lower	upper	ESS
β_0	1.902	1.901	1.745	2.062	3598.411
$\boldsymbol{\beta}_1$: (1) Binding	-0.073	-0.073	-0.203	0.054	15715.895
$\boldsymbol{\beta}_2$: (2) No-Change	-0.439	-0.438	-0.548	-0.327	9320.829
$\boldsymbol{\beta}_3$: (3) SS6	-0.676	-0.676	-0.768	-0.583	10195.538
$\boldsymbol{\beta}_4$: (4) Older Group	-0.411	-0.410	-0.570	-0.244	3357.847
$\boldsymbol{\beta}_5$: 1×3	0.253	0.252	0.123	0.382	16018.744
$\boldsymbol{\beta}_6$: 2×3	-0.268	-0.268	-0.380	-0.158	10051.656
$\boldsymbol{\beta}_7$: 1×4	-0.063	-0.063	-0.190	0.068	15641.582
$\boldsymbol{\beta}_8$: 2×4	0.224	0.224	0.114	0.334	9937.636
$\boldsymbol{\beta}_9$: 3×4	0.019	0.019	-0.072	0.113	10418.598
$\boldsymbol{\beta}_{10}$: $1 \times 3 \times 4$	0.059	0.059	-0.072	0.186	16339.511
$\boldsymbol{\beta}_{11}$: $2 \times 3 \times 4$	-0.038	-0.038	-0.149	0.072	9504.982
σ_s	0.449	0.444	0.318	0.591	7451.405

Note: The effects coded variables were as follows: (1) Binding = 1, No-Change = 0, Colour = -1, (2) Binding = 0, No-Change = 1, Colour = -1, (3) SS3 = -1, SS6 = 1, (4) Younger = -1, Older = 1. Interaction contrasts were products of these effects coded variables.

winning model *did not* include the age × condition interaction. Rather this model is comprised of the three main effects plus an interaction between age and set size. A contrast of models 1 and 3 revealed good evidence against the age × condition interaction ($B_{1,3} = 4.529$). Further there was no suggestion that this was modulated by set size, ($B_{5,6} = 2.078$). By contrast there was strong evidence for the age-group interaction with set size, ($B_{1,4} = 15.17$), confirming the reliability of the pattern shown in Figure 5.7.

In order to test whether the omission of trials involving a change to location had an effect on sensitivity we followed this up by comparing performance in Experiment 8 to the data from the mixed condition of Experiment 7. To compare the two we removed the location condition data from Experiment 7 and conducted a BANOVA with the additional factor of Experiment. First we assessed whether overall performance differed between the two experiments. There was no evidence for this suggestion as a model omitting the effect of Experiment was favoured by over 2-to-1. There was substantial evidence against the age × condition interaction,

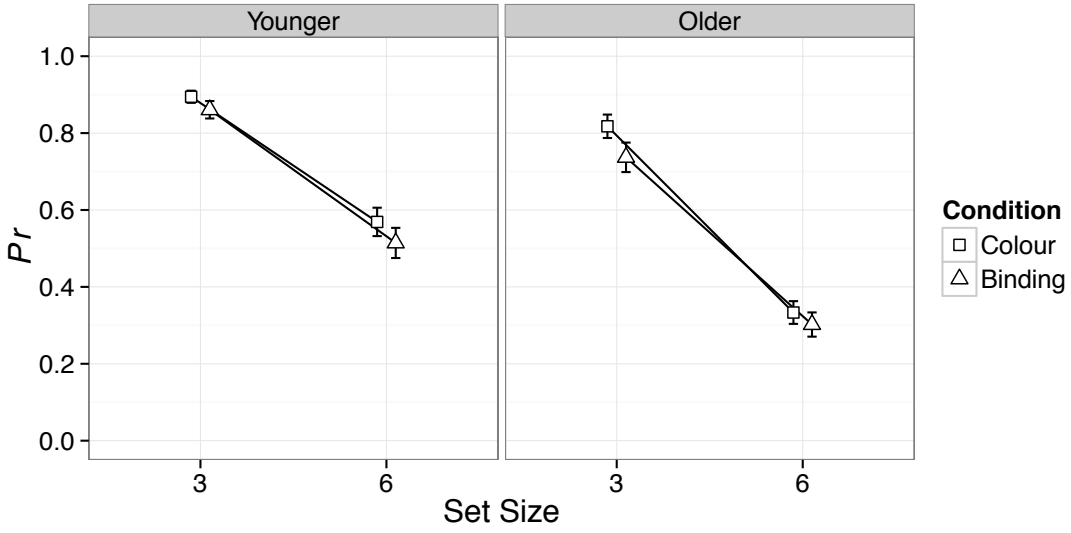

Figure 5.7: P_r (corrected recognition) across age groups and experimental conditions in Experiment 8. Error bars are ± standard error.

Table 5.10: Log Bayes factors for analysis of sensitivity (P_r) in Experiment 8

	Model	$\log(B_{M,0})$	% error
1	$P_r \sim$ SS + C + AG + SS:AG + ID	124.40	1.58
2	$P_r \sim$ SS + C + SS:C + AG + SS:AG + ID	122.90	2.47
3	$P_r \sim$ SS + C + AG + SS:AG + C:AG + ID	122.89	0.87
4	$P_r \sim$ SS + AG + SS:AG + ID	121.68	1.61
5	$P_r \sim$ SS + C + SS:C + AG + SS:AG + C:AG + ID	121.43	1.40
6	$P_r \sim$ SS + C + SS:C + AG + SS:AG + C:AG + SS:C:AG + ID	120.70	2.15
7	$P_r \sim$ SS + C + AG + ID	119.30	0.53
8	$P_r \sim$ SS + C + SS:C + AG + ID	117.85	1.74
9	$P_r \sim$ SS + C + AG + C:AG + ID	117.82	0.79
10	$P_r \sim$ SS + AG + ID	117.02	0.49
11	$P_r \sim$ SS + C + SS:C + AG + C:AG + ID	116.37	1.81
12	$P_r \sim$ SS + C + ID	112.87	0.57
13	$P_r \sim$ SS + C + SS:C + ID	111.42	1.40
14	$P_r \sim$ SS + ID	110.55	0.76
15	$P_r \sim$ AG + ID	4.08	0.34
16	$P_r \sim$ C + AG + ID	3.17	0.88
17	$P_r \sim$ C + AG + C:AG + ID	1.62	0.70
18	$P_r \sim$ C + ID	-0.95	0.33

Note: All Bayes factors are relative to a null model with a random participant effect only (model 0: $P_r \sim$ ID). AG = Age-Group, C = Condition, SS = Set Size, ID = participant ID, and ':' denotes an interaction effect

($B_{R,F} = 6.296$), as well as the suggestion that this interaction differed across the two Experiments, ($B_{R,F} = 4.799$). Thus, we have fairly clear evidence that omitting trials in which a change occurred only to location *did not* lead to the emergence of an age-related binding deficit in terms of sensitivity.

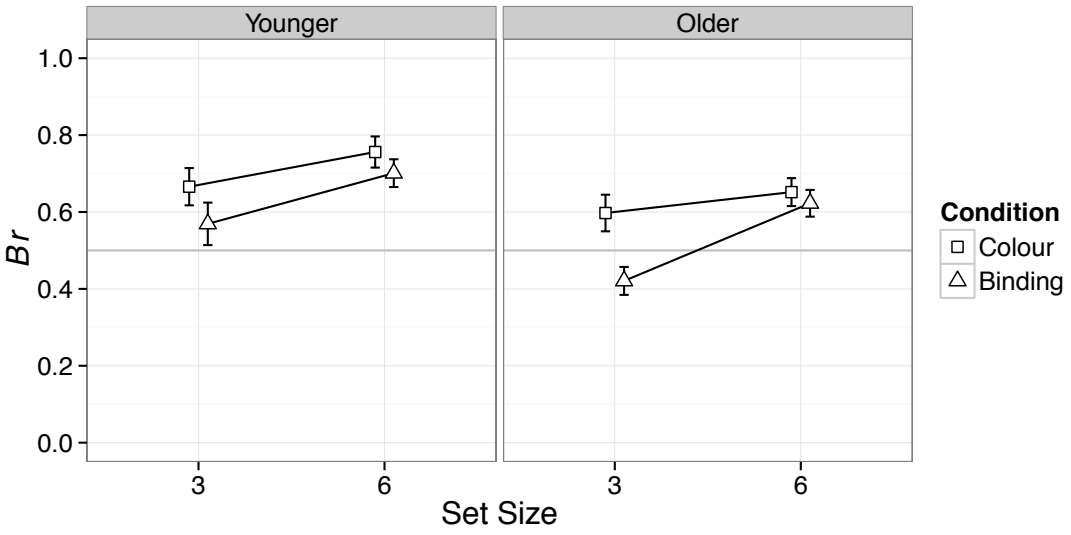

Figure 5.8: B_r across age groups and experimental conditions in Experiment 8. Error bars are ± standard error.

Bias

Estimated guessing bias (B_r) across experimental conditions and age-groups is presented in Figure 5.8. The Bayesian ANOVA (Table 5.11) revealed that the winning model included all three main effects with an interaction between memory condition and set size. Testing the strength of evidence for the omission of the key age by condition interaction we can compare the winning model to model 5. This revealed that the omission of the interaction is favoured over its inclusion by 4-to-1 ($B_{1,5} = 4.019$). Figure 5.8 gives some reason to believe that age differences in bias are more pronounced for binding at set size 3. However, the data do not clearly adjudicate on this matter as seen in the contrast of model 9, which omits the three way interaction, to model 11, which includes it ($B_{9,11} = 2.076$). Finally, while the winning model included the condition × set size interaction we see that this model was marginally favoured over model 2 which left this component out ($B_{1,2} = 1.117$). Thus, while Figure 5.8 gives the impression that set size had a slightly larger effect on bias in the binding condition, the evidence for this interaction is equivocal.

As before we conducted an analysis combining the current data set with bias estimates from the mixed condition of Experiment 7 (omitting the data from the location condition). There was no suggestion that, overall, guessing bias differed

Table 5.11: Log Bayes factors for analysis of sensitivity (B_r) in Experiment 8

	Model	$\log(B_{M,0})$	% error
1	$B_r \sim$ SS + C + SS:C + AG + ID	12.88	0.78
2	$B_r \sim$ SS + C + AG + ID	12.77	0.83
3	$B_r \sim$ SS + C + SS:C + ID	12.02	1.21
4	$B_r \sim$ SS + C + ID	11.92	0.31
5	$B_r \sim$ SS + C + SS:C + AG + C:AG + ID	11.49	2.14
6	$B_r \sim$ SS + C + SS:C + AG + SS:AG + ID	11.39	1.23
7	$B_r \sim$ SS + C + AG + C:AG + ID	11.35	0.79
8	$B_r \sim$ SS + C + AG + SS:AG + ID	11.28	1.20
9	$B_r \sim$ SS + C + SS:C + AG + SS:AG + C:AG + ID	9.97	1.10
10	$B_r \sim$ SS + C + AG + SS:AG + C:AG + ID	9.87	1.44
11	$B_r \sim$ SS + C + SS:C + AG + SS:AG + C:AG + SS:C:AG + ID	9.24	2.86
12	$B_r \sim$ SS + AG + ID	8.65	0.32
13	$B_r \sim$ SS + ID	7.80	0.26
14	$B_r \sim$ SS + AG + SS:AG + ID	7.16	1.74
15	$B_r \sim$ C + AG + ID	4.15	0.47
16	$B_r \sim$ C + ID	3.34	0.38
17	$B_r \sim$ C + AG + C:AG + ID	2.71	0.51
18	$B_r \sim$ AG + ID	0.79	0.78

Note: All Bayes factors are relative to a null model with a random participant effect only (model 0: $B_r \sim$ ID). AG = Age-Group, C = Condition, SS = Set Size, ID = participant ID, and ':' denotes an interaction effect

between the two experiments as a model omitting the main effect of Experiment was more likely given the data ($B_{R,F} = 4.669$). Further there was substantial evidence against interactions between age-group and experiment ($B_{R,F} = 4.258$) and age-group and memory condition ($B_{R,F} = 4.086$). Finally, a model omitting the three-way interaction between age, condition, and experiment was almost 5 times more likely given the data than the full model, ($B_{R,F} = 4.95$). It seems, then, that omitting the location change trials from the experimental design had no important effect on response bias, in line with our analysis of sensitivity above.

Discussion

Contrary to our expectation, bringing our design closer to that of Cowan et al. (2006) did not allow us to recreate their finding of a specific age related binding deficit. It is likely that another methodological difference between the two studies can account for the discrepant findings. For example, as discussed above, Cowan et al. (2006) presented a whole display at test with a single circled item whereas we presented only one item in the test array. Read et al. (2016), who also failed to demonstrate an age-related colour-location binding deficit, suggested that the whole display used

by Cowan et al. may have caused disproportionate binding specific interference at test (cf. Wheeler & Treisman, 2002). Given the evidence we reported in Chapter 2 we do not think that at-test interference can account for Cowan and colleague's findings. In that Chapter we demonstrated that there is very little performance cost associated with probing change detection with a whole display relative to a single probe. Previous demonstrations of whole display interference (e.g. Wheeler & Treisman, 2002) arose due to a lack of a principled measure to compare the two tasks which place different demands on VWM capacity (see also, Cowan et al., 2013).

Rather we suspect it was the inclusion of duplicates in the memory and test arrays of Cowan et al. (2006) that was the main contributing factor to their pattern of results. In the Introduction we noted that the probed item in the colour condition always was unique whereas in the binding condition it was always a duplicate (see Cowan et al., 2014). In the blocked condition participants knew exactly which kind of change to expect and thus may have appreciated that in the binding condition, for example, it was sufficient to note only the repeated colours. Age differences in apprehending this intricate aspect of the task would contribute to age differences in task performance. Further, it seems likely that noting which colours are duplicated in an array is a more simple task than noting which colours are unique. This would introduce a benefit for binding trials especially when trials were blocked. This is evident in Cowan et al's data (see Table 2 on page 1095) as there was a clear difference in terms of sensitivity (d') between the colour and binding conditions in Experiment 1A in which these trials were mixed. In Experiment 2A, in which trials were blocked, there was very little difference between the two conditions and, in fact, for younger adults sensitivity to binding changes was greater than that for colour changes. This suggests that participants were able to make use of additional information in the probe array to guide detection of binding changes when they were aware that this was the type of change to expect.

It is clear that the use of a single probe without the presence of unprobed items is a better way of addressing the question of the efficacy of feature binding in healthy ageing (see also, Read et al., 2016, for similar findings). If there was a true effect

of mixing different trial types on the sensitivity of older adults to binding changes the present paradigm would have shown this. On the contrary, three experiments (Chapter 4 and the present experiments) have demonstrated no effect of mixing versus blocking trials.

5.4 General Discussion

Early investigations of healthy ageing and feature binding in VWM assessed the ability to retain the correspondence between object identity and location. This work led to the suggestion that healthy adult ageing is accompanied by a specific deficit in retaining what was where *over and above* the ability to retain what *or* where individually (Mitchell, Johnson, Raye, Mather, & D'Esposito, 2000; Mitchell, Johnson, Raye, & D'Esposito, 2000). Follow up studies have largely supported these initial suggestions (Borg et al., 2011; Cowan et al., 2006; Fandakova et al., 2014; Peich et al., 2013), although the weight of evidence offered by these studies is questionable. Also, there have been some reports that do not show this (Bopp & Verhaeghen, 2009; Olson et al., 2004; Pertzov et al., 2015; Read et al., 2016). The present study falls into the latter category and was able to go further than previous work, via the use of Bayesian statistical methods, to quantify the evidence *against* a differential effect of age on VWM for item-location bindings. In Experiment 7 the sensitivity data were over 17 times more likely under a model omitting the age-group × condition interaction and there was strong evidence (over 10-to-1) against modulation by block type; this was also the case in our analyses of response bias.

At this point it is useful to review previously offered explanations for failures to detect an age-related location binding deficit. Olson et al. (2004) found evidence that older adults bound items in different locations into a configural representation in a similar manner to younger adults. Both groups' ability to detect changes to location was disrupted by changes to non-probed (contextual) items. They suggested that, as they had used shorter retention intervals (around 1.5 seconds) relative to the earlier studies of Mitchell and colleagues (8.5 seconds), that requirement to retain items for extended periods of time may lead to the emergence of specific deficit. Recently,

Pertzov et al. (2015) directly investigated whether temporary memory for object-location associations deteriorates over a longer retention interval. In their study participants of various ages, ranging from 19 to 83, were required to remember 3 difficult-to-name fractals over a 1 or 4 second interval. Following this interval participants had to select a previously seen fractal from a choice of two and then drag the object to its remembered location, thus giving the *precision* with which location was retained. The main finding was that older adults were more likely to produce swap errors, in which a fractal was wrongly located to a location occupied by another stimulus. However the frequency of swap errors was no greater than would be predicted from older adults' poorer VWM for object identity alone. Crucially the retention interval had no effect on the likelihood of swap errors. While the intervals used by Pertzov and colleagues were not as long as the initial delays used by Mitchell et al. these findings do not support a role of retention interval in the emergence of a binding deficit.

Another feature shared by previous reports arguing for an age-related location binding deficit is the sequential presentation of memory items, as opposed to the simultaneous presentation of an array used here. It has been suggested that sequential presentation of memoranda supports encoding that relates each item to an external frame of reference (and is thus item specific) whereas simultaneous presentation allows items to be encoded as part of a global configuration (Jiang et al., 2000; Lecerf & De Ribaupierre, 2005). This overlaps greatly with the distinctions discussed in Chapter 1, such as that between categorical representation of the relationship between objects (such as blue is above red), that may be encouraged by simultaneous presentation, versus more precise coordinate level representations, that may be called upon more with sequential presentation (Baddeley, Jarrold, & Vargha-Khadem, 2011; Postma et al., 2008). Further, age differences in temporary memory for locations have been shown to be larger when the task requires recall of sequentially presented locations relative to a simultaneous array (Oosterman et al., 2011). Therefore, it is possible that previous reports of age-related binding deficits for location have been driven by the use of sequential presentation, which may sup-

port a more veridical binding between object and location rather than a configural representation. Thankfully, Read et al. (2016) recently addressed this distinction in their change detection experiments by directly comparing the two modes of presentation. They found that, overall, detecting binding changes was more difficult when the colour-location pairs had been presented sequentially, however, this was true for both age-groups. There was no evidence for a disproportionate effect of age on binding change detection in both presentation conditions.

Finally, another noted difference between those who claim to find age-related difficulty in retaining what went where and those that do not, is the use of easy-to-name material. Mitchell and colleagues, for example, used clip-art-like images of familiar items presented in a 3×3 grid, whereas Fandakova et al. (2014) used letters presented in one of six boxes arranged horizontally. One possibility suggested by Pertzov et al. (2015) is that younger adults are more likely to engage in verbal strategies that support the retention of associations (cf. Stefurak & Boynton, 1986). When stimuli are difficult to verbalise younger adults do not get this 'boost' in the conjunction condition and no interaction is observed. One may also suggest that, rather than older adults being less likely to adopt verbal strategies, both groups may attempt to name stimuli making the task increasingly *associative* or *relational* in nature (i.e. retaining *book + top-left*). As outlined in the first Chapter there is a well known associative deficit with healthy ageing (T. Chen & Naveh-Benjamin, 2012; Old & Naveh-Benjamin, 2008a) thus the use of easy to verbalise material without articulatory suppression may lead to the appearance of a specific deficit (Peterson & Naveh-Benjamin, 2016). This distinction between relational and conjunctive binding has gained increasing research interest recently (e.g. Ecker et al., 2013; Parra, Fabi, et al., 2015; Piekema et al., 2010), however, to our knowledge only one study has directly assessed this distinction in the context of healthy ageing (van Geldorp et al., 2015). In Chapter 6 we report a small study contrasting colour-shape binding when these features are either present as part of separate objects (colour and shape are extrinsic) or as intrinsic features of the same object.

These factors may prove crucial, however, as discussed in detail in the Intro-

duction, the evidence in favour of age-related location binding deficits is highly questionable. Many previous studies have not reported sufficient evidence to support a specific age-related location binding deficit (Fandakova et al., 2014; Mitchell, Johnson, Raye, Mather, & D'Esposito, 2000; Mitchell, Johnson, Raye, & D'Esposito, 2000) or contain errors that may invalidate crucial results (Borg et al., 2011). The present work suggests that, for simple stimuli retained over a very brief period, a more parsimonious explanation of older adults' change detection performance in a general decline of VWM with some evidence for more impaired change detection relative to sameness detection. This general decline of VWM with age has been noted many times before and explanations of this tend to fall into capacity based and attentional control based accounts. In the next Chapter we use raw change detection accuracy from Experiment 7 reported here and the experiment reported in Chapter 4 to explore the viability of these explanations using processing models from the slots conception of VWM.

Chapter 6

The Effect of Age on Intrinsic and Extrinsic Binding

6.1 Introduction

The evidence provided in the previous three chapters against a specific object feature binding deficit in healthy older adults stands in stark contrast to the associative LTM deficit (Old & Naveh-Benjamin, 2008a; Spencer & Raz, 1995). Recent work suggests this is not only a LTM phenomenon and that the associative deficit is present in short-term/ working memory (T. Chen & Naveh-Benjamin, 2012; see also, Hartman & Warren, 2005). As noted by T. Chen and Naveh-Benjamin (2012) there are two clear, somewhat overlapping, differences between studies that find working memory binding deficits and those that do not. The former tend to use more complex, ecologically valid stimuli (e.g. pictures of faces and scenes) whereas the latter use simple colours and geometric shapes. Given that complex stimuli inherently contain more features that presumably need to be bound together, it may be this increased binding load that leads to the age-related associative deficit. Alternatively these two literatures could be assessing fundamentally different forms of binding with older adults exhibiting a deficit on tasks requiring between item binding (*extrinsic* binding) and not on tasks where features must be combined within objects (*intrinsic* binding). This distinction is discussed in more detail below. On the basis of the

current literature the influence of these two factors is hard to gauge, therefore the present experiment aimed to directly contrast these theoretically different forms of binding using identical stimulus features.

Binding Intrinsic and Extrinsic Features

A crucial distinction has often been made in the LTM literature between memories for features intrinsic or extrinsic to an object or event (e.g., Baddeley, 1982). Intrinsic features are the defining characteristics of an item, whereas extrinsic features are those that provide contextual detail. Formation and retention of these memories is proposed to rely on fundamentally different binding mechanisms (Zimmer & Ecker, 2010; Zimmer et al., 2006). According to the type-token model (Zimmer & Ecker, 2010), features intrinsic to an object or event are proposed to be integrated relatively automatically into an *object token* when attention is directed to them (see also, Duncan, 1984; Treisman, 2006). On the other hand memory for contextual features accompanying, but extrinsic to, an object are proposed to be more effortfully integrated into an *episodic token*. What is considered intrinsic and extrinsic is difficult to strictly define and will surely depend on the expectations and goals of the observer (e.g., Marr, 1982). Nevertheless, there have been some experimental investigations that have aimed to contrast these theoretically different representational structures by presenting different features within integrated objects or as spatially distinct items that must be associated. These investigations have repeatedly shown that associative recall is better when features are presented as an integrated whole as compared to two separate entities (e.g., Arnold & Bower, 1972; Asch, Ceraso, & Heimer, 1960; Ceraso, Kourtzi, & Ray, 1998; Walker & Cuthbert, 1998; Wilton, 1989).

Recent work using the change detection paradigm has also highlighted the importance of the intrinsic/ extrinsic distinction. Ecker et al. (2013) showed that when colour-shape conjunctions were presented as intrinsic relations, with the colour filling the shape, there was evidence that task irrelevant changes to colour affected change detection performance for shape, suggesting the features were obligatorily

bound together. However, when the colour was presented as the background, that is extrinsic to the shape, there was no evidence of this obligatory binding. Further studies assessing intentional binding using similar methodology have shown that detecting changes to conjunctions of features is much better when stimuli are presented as unitary objects relative to when one feature acts as context (background) for the other (Delvenne & Bruyer, 2004; Xu, 2002).

Therefore, the available body of evidence suggests that features intrinsic to an object are bound automatically at encoding (see also, e.g., R. J. Allen et al., 2006; C. C. Morey & Bieler, 2013), whereas binding extrinsic features (e.g. background) to an object appears to come at more of a cognitive cost. Given that ageing disproportionately affects performance on memory tasks requiring controlled, effortful processing but leaves automatic processes largely intact (Craik & Bialystok, 2006; Jennings & Jacoby, 1993) one would expect the pattern of results obtained so far, with age-related WM binding deficits for between item associations (e.g., T. Chen & Naveh-Benjamin, 2012) but not associations within items (e.g., Brown et al., 2016).

There have been few direct comparisons of the different forms of memorial binding in the context of ageing research, however, several studies in the LTM associative deficit literature point towards the importance of the intrinsic/ extrinsic distinction. It has been repeatedly demonstrated that the associative deficit can be reduced, if not eliminated, by requiring memory for semantically related units of information (i.e. relations that can be processed as a single chunk, Badham, Estes, & Maylor, 2012; Naveh-Benjamin et al., 2003; Naveh-Benjamin, Craik, Guez, & Kreuger, 2005). However, the effect of unitisation does not depend on pre-existing semantic relations and can be produced with simple encoding manipulations. This was recently shown by Bastin et al. (2013), who in two sessions required participants to learn associations between words and their corresponding background colour (red or green) for a subsequent source-recall test. Crucially in the different sessions, participants were either encouraged to encode the word-colour pairings as unitised items (imagine the item filled in the colour), or to encode the colour as a contextual detail to be associated with the word (imagine the item interacting with another item

filled in the background colour). Despite participants having to remember identical information across these sessions, the effect of age on source-recall was significantly reduced by instructions to encode the pairings as a unit.

The importance of the intrinsic/ extrinsic distinction goes beyond the ease with which associations can be formed as these representational structures are proposed to map onto the phenomenological sensations of *recollection* and *familiarity*. For example, the type-token model of Zimmer and Ecker (2010) suggests that the extrinsic features accompanying episodic tokens facilitate feelings of recollection by reinstating spatiotemporal context, whereas reinstatement of an object token supports familiarity only—a feeling of knowing that the object has previously been encountered. Much of the evidence for this overlap comes from the observation that both extrinsic binding and recollection appear to rely on the hippocampus, whereas intrinsic binding and familiarity appear to rely on parahippocamal areas, such as the perirhinal cortex (Zimmer & Ecker, 2010; Ranganath, 2010).

Older adults' memorial experience appears to be impoverished for contextual detail, making their memory judgements more likely to be based of feelings of 'knowing' that the item has previously been encountered rather than 'remembering' (Mäntylä, 1993). Studies using the more objective process dissociation paradigm have shown that older adults' memory judgements are more likely to be based on an automatic feeling of familiarity rather than recollection of the specific context under which an item was encountered (Jennings & Jacoby, 1993; see also, M. G. Rhodes et al., 2008). Thus if the increased reliance on feeling of familiarity seen with healthy ageing is due to better preserved intrinsic binding we may expect less of a binding cost when features are presented as a conjunction. Interestingly a recent meta-analysis suggests that preserved familiarity based processing may be a hallmark of healthy ageing, whereas patients with Alzheimer's disease exhibit a pronounced deficit in both recollection and familiarity (Koen & Yonelinas, 2014; see also, Tse, Balota, Moynan, Duchek, & Jacoby, 2010).

As alluded to above, there is growing neuropsychological and neuroimaging evidence for these different levels of binding. In the LTM literature it has been sug-

gested that extra-hippocampal regions, such as the perirhinal cortex, are responsible for forming object representations, whereas the hippocampus binds these object representations to their spatio-temporal context (see, Davachi, 2006; Diana, Yonelinas, & Ranganath, 2007, for reviews). This mapping shows considerable overlap with regions associated with familiarity and recollection, respectively (Eichenbaum, Yonelinas, & Ranganath, 2007). While it has been argued that hippocampal involvement is a hallmark of LTM processes (Baddeley, Jarrold, & Vargha-Khadem, 2011) there is increasing evidence that this area plays a role in relational binding in WM. In two fMRI studies Piekema and colleagues required participants to retain various different kinds of association in WM for a recognition probe (Piekema, Kessels, Rijpkema, & Fernández, 2009; Piekema et al., 2010). When participants were maintaining the association between images of faces and houses activity was observed in the MTL including the hippocampus, parahippocampal gyrus and amygdala. However, when participants in this study were retaining an intrinsic association (a face in a certain colour), no such MTL activity was observed (see also, Parra, Della Sala, Logie, & Morcom, 2014, for evidence with coloured shapes).

In line with this, Parra, Fabi, et al. (2015) report the case of patient AE, an individual with MTL damage (including the right hippocampus) as a result of stroke, performing tasks requiring the recall of object-colour associations from WM. When object and colour were unitised, AE was as good as age-matched controls at associative recall but when the features were spatially distinct he exhibited a pronounced deficit. On the other hand, Baddeley, Allen, and Vargha-Khadem (2010) report the case of Jon, an amnesic patient with bilateral hippocampal lesions, who was able to detect changes to shape-colour binding as efficiently (if not more efficiently) as control participants. Crucially, presenting the to-be-associated features as spatially distinct, extrinsic items did not affect Jon's performance. Thus while patient AE's pattern of performance is in line with neuroimaging findings, neuropsychological findings are mixed. There are a number of important differences between these studies that may account for the divergent findings. Namely, the origin of the hippocampal damage experienced by the two cases (Jon's amnesia is congenital

whereas AE's is acquired) and the nature of the tasks they were asked to perform (recall versus recognition). Larger studies of amnesic patients will be necessary to address these potential mediating factors.

While the retention of intrinsic bindings in WM does not appear to require hippocampal involvement the neural underpinnings of this function are less clear. Some fMRI studies assessing change detection with multi-featured objects have found greater activity in the intraparietal sulcus (IPS) when participants are retaining conjunctions of features relative to individual features (Song & Jiang, 2006; Xu, 2007; Xu & Chun, 2006). Others, however, have not found conjunction related activity in the IPS, instead finding more widely spread parietal activity (Parra et al., 2014). All of these studies have found increased activity in the lateral occipital complex associated with storing intrinsic features in WM. Thus it appears that temporary retention of feature conjunctions occurs at an earlier stage of the processing hierarchy than associative (hippocampal) binding. It is well known that healthy adult ageing is associated with progressive loss of MTL volume, particularly the hippocampi (Raz & Rodrigue, 2006), and this is considered an important contributor to the associative deficit (Shing et al., 2010). On the other hand, the regions associated with intrinsic binding, thus far, appear to be relatively spared by normal ageing (Raz & Rodrigue, 2006).

Therefore, it seems probable that the mixed results regarding age-related working memory binding deficits have largely been caused by the assessment of different binding mechanisms. However, the possibility remains that stimulus complexity (real world scenes and faces versus abstract polygons and colours) underlies the difference (cf. T. Chen & Naveh-Benjamin, 2012). Therefore, the present experiment addresses this directly by comparing younger and older adults' ability to bind colour and shape when these features are presented as integrated objects as opposed to when they are presented as two distinct items. Given that WM binding deficits have been observed with real-world stimuli (T. Chen & Naveh-Benjamin, 2012) if we do not find an extrinsic binding deficit in the present circumstance this may suggest that older adults do not struggle to bind simple features, regardless of how

they are presented. If, on the other hand, older adults find retaining the feature association particularly difficult when they are presented extrinsically this would point to a potential resolution of previous findings.

This is not the first study to address this question; recently, van Geldorp et al. (2015) have directly compared intrinsic and extrinsic binding using colour-shape stimuli (identical to those used in Chapters 4 and 5). In their conjunctive (intrinsic) condition the colours and shapes appeared as part of the same object and in their relational (extrinsic) condition colours and shapes appeared as spatially distinct objects linked by a single line. They found older adults were less able to reconstruct pairings of shape and colour but the manner of presentation *did not* modulate the age effect. This goes against what we may predict on the basis of the literature outlined above. However, it is important to note that this study did not include independent measures of item and associative memory, making conclusions regarding binding deficits difficult. Further, as the features in the relational condition were spatially separate this doubled the number of to-be-attended locations at encoding which could conceivably have affected item memory (Xu, 2002). In the present experiment we avoided the confound between the type of relation (intrinsic/ extrinsic) and the number of spatial locations by presenting items sequentially. Each memory display included an abstract polygon and a circle (see Figure 6.1). The to-be-remembered colour was either presented in the to-be-retained polygon (intrinsic) or in the otherwise irrelevant circle (extrinsic). These presentation formats were mixed together to ensure that observers always had to attend to both array components—the crucial difference was whether the features came from the same source or distinct items.

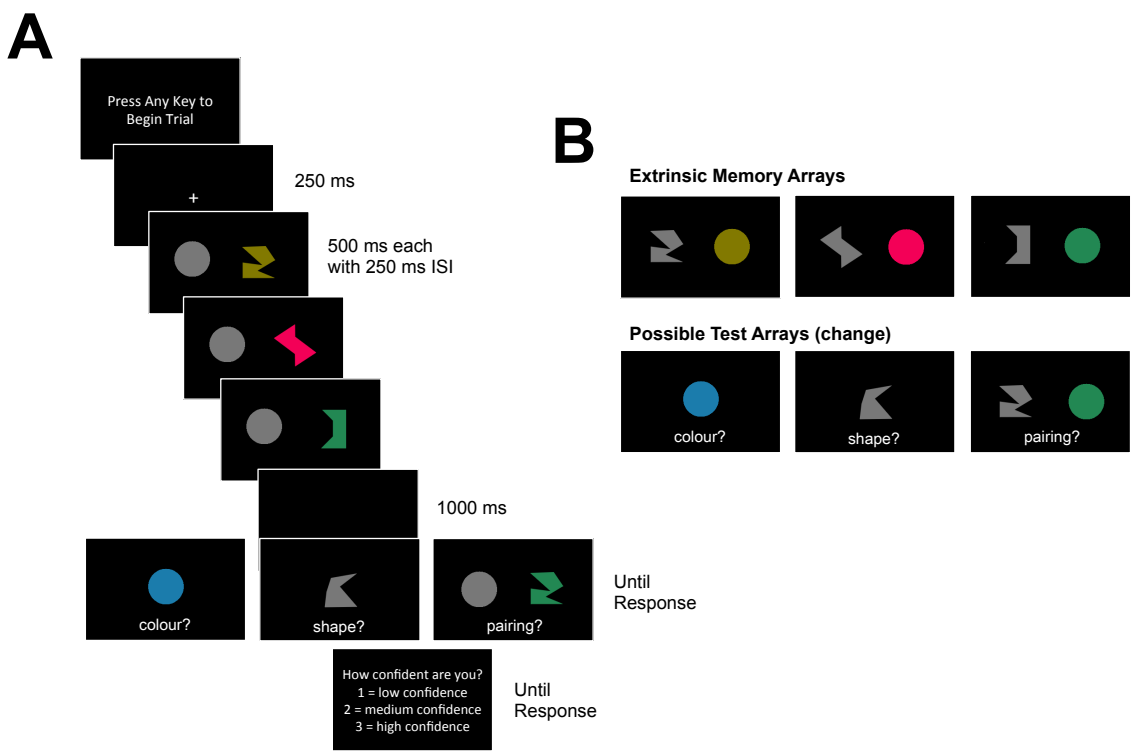

Figure 6.1: (A) The general trial procedure used in Experiment 9 depicting an example where colour and shape are presented as *intrinsic* features. (B) An example of memory arrays and possible test arrays for a trial in which colour and shape are *extrinsic*. *Note:* Figure is not drawn to scale and all test arrays depict change trials.

6.2 Experiment 9 – Intrinsic versus Extrinsic Change Detection

Method

Participants

Twenty-four younger and 24 older adults took part in this experiment. All of the younger group and 23 of the older group also took part in the second experiment reported in Chapter 5 prior to completing this study, making the session approximately 45 minutes long. One older adult from Chapter 5 was unable to finish both sections of the session within the allotted time and they were replaced by an individual from the same volunteer pool who completed only this study.

Stimuli

In assessing the distinction between intrinsic and extrinsic binding in VWM it is important to ensure that stimuli are difficult to encode verbally, as verbal encoding has been shown to blur this distinction (Walker & Cuthbert, 1998). To this end we used difficult to name shapes and a wide variety of discriminable colours. The shapes were selected without replacement from a set of 16 abstract polygons with 6–8 angles (taken from, Parra, Abrahams, Logie, & Della Sala, 2010; Ecker et al., 2013). The colours were taken from a set of 360 surrounding a colour wheel in the CIE L*a*b* colour space (L* = 50, centered at a* = b* = 20, radius = 60). This colour space is widely used in studies assessing delayed reproduction from VWM (e.g. van den Berg, Shin, Chou, George, & Ma, 2012; W. Zhang & Luck, 2008) and allows us more control over the discriminability of colours. To make sure that colours were discriminable within-trials a set of 6 colours was selected at the beginning of each trial, each separated by 60° on the colour wheel, to act as the trial colour set. This separation was selected to be highly discriminable for both younger and older adults (Peich et al., 2013).

Selecting from such a large set of equiluminant colours was also important given that in some of our previous studies (Chapters 4 and 5) participants reported becoming familiar with certain combinations of features or adopting strategies to remember colours (e.g. focus on 'bright' ones). On each trial 3 colours were selected from the set of 6 and presented during the memory sequence. For trials on which a colour change occurred one of the remaining 3 colours was selected and presented at test. Each study screen contained one abstract polygon (selected randomly without replacement) and a circle. Each item in the study/test screens subtended approximately 2° by 2° and were separated centre-to-centre by 4° at an approximate viewing distance of 50 cm. On each trial it was determined at random whether the abstract shape appeared to the left or right of the circle and this remained consistent within the trial. Stimuli were presented against a black background on an 18" E96f+SB ViewSonic monitor with a resolution of 1024 × 768 and a refresh rate of 100 Hz.

In the *intrinsic presentation* condition the colour and shape appeared together,

with the colour filling the polygon, accompanied by the task irrelevant circle filled in grey (L* = 50, a* = b* = 0; see Figure 6.1A). In the *extrinsic presentation* condition the to-be-remembered shape was presented filled in grey next to the circle which was filled in the to-be-remembered colour (see Figure 6.1B).

The nature of the test screen differed depending on whether it was probing memory for colour only, shape only, or colour and shape pairing. A test screen probing memory for colour consisted of a single coloured circle at the centre of the screen whereas a test screen probing memory for shape contained a single shape filled in grey. The test screen probing memory for the exact pairing of colour and shape consisted of an abstract shape and circle on either side of the centre of the screen and the manner of presentation matched the memory arrays of that trial (e.g., if the shape and colour were presented in separate items, extrinsic presentation, the test array also presented the features in this way). Feature changes (shape or colour) involved presenting a brand new feature that did not appear in the initial memory array sequence. A pairing (binding) change involved a repairing of features that appeared in separate memory arrays (see Figure 6.1 for examples of these changes). As these trials were mixed together the test screens also contained a text prompt ('colour?', 'shape?', or 'pairing?') to guide participants' responses.

Design and Procedure

The general trial procedure is presented in Figure 6.1A. Participants began each trial with a keypress which was followed by a 250 ms fixation screen. Three memory screens (containing an abstract polygon and a circle) were then presented sequentially for 500 ms each with a 250 ms blank inter-stimulus-interval. After the last memory screen was presented there was a 1000 ms retention interval followed by the presentation of the test array which probed memory for the colours, the shapes, or the colour-shape pairings presented. These different trial types were mixed together so participants were unaware of what aspect of the stimuli would be probed. Participants were required to indicate whether the probe was the *same* as one of the to-be-remembered sequence or was *different* by using the z and m keys, respectively.

Following the response participants indicated how confident they were in their response from 1 (low confidence) to 3 (high confidence). As will become clear below, collecting confidence ratings allows us to construct receiver operating characteristic (ROC) curves (Yonelinas & Parks, 2007) and conduct a truly non-parametric assessment of sensitivity differences. The different presentation formats were also mixed within the same trial-blocks to ensure that participants attended to both of the elements in the memory displays.

Prior to completing the task detailed instructions were given emphasising that the colour and the shape were the to-be-remembered features and participants would be asked about the colour, shape and pairing equally often. Visual examples of the kinds of change to expect were also given. Participants were given 12 practice trails with one same and one different trial for each presentation and test type. For the main part of the experiment participants completed 108 experimental trials overall (breaks were offered after 36 and 72 trials) with 18 trials for each combination of presentation and test type (half same, half different).

Analysis

The use of a confidence rating procedure allows us to construct empirical ROC curves reflecting false-alarm and hit rates at different levels of confidence. In our previous experiments, and in previous work on feature binding and healthy ageing, we have only had a single hit and false-alarm rate pair per condition to estimate sensitivity or a similar metric (d', P_r, k). As shown in Chapter 8, if the assumptions underlying these measures are invalid, researchers are at risk of erroneously detecting an interaction effect. The additional information provided by the evaluation of confidence following each *same* or *different* response can provide one with greater confidence in their assessment of sensitivity.

Green (1964) showed that the area under the ROC curve is equal to the expected proportion correct of an unbiased observer given a two-alternative forced choice. That is, area provides an estimate of an observer's ability to choose between a target and lure without recourse to assumptions regarding the underlying recognition

process. One can attempt to estimate area with a single hit and false-alarm pair, but these attempts have been far from satisfactory (see Chapter 8). A 3 point rating scale for *same* and *different* responses produces a 5 point empirical ROC curve (final point must be (1, 1)) from which area can be estimated. To do this we use a more conventional definition of hits and false-alarms, where a hit is correct identification of a previously encountered item (i.e. a correct *same* response) and a false-alarm incorrectly identifies an item that was not encountered (i.e. an incorrect *same* response). A full explanation of how the empirical ROC is constructed can be found in Appendix B.

We calculated two measures of area under the curve; the first, referred to as A_g (Pollack & Hsieh, 1969), estimates area by tracing a line from each (f, h) pair to $(f, 0)$ and summing up the area of the resulting trapezoids. This measure provides a very conservative estimate of the area under the curve (Macmillan & Creelman, 2005; Stanislaw & Todorov, 1999) so in addition we considered the measure, A_z (Swets, 1988). A_z is derived from SDT with underlying Gaussian evidence distributions but it does not make the equal variance assumption, and can thus accommodate asymmetrical ROC curves. A_g is purely a geometric estimate of area and makes no reference (explicit or implicit) to underlying evidence distributions. Therefore, unlike A', A_g can truly be considered a non-parametric estimate of sensitivity uncontaminated by response bias and consequently we focus our results section on this measure. The exact method of calculation used for these measures can also be found in Appendix B.

Results

The empirical ROC curves for each age-group across the different presentation and test types are presented in Figure 6.2 and Figure 6.3 presents the estimate of area A_g. As in previous Chapters, to assess the evidence for and against main- and interaction-effects of interest we conducted a Bayesian ANOVA in which models of varying complexity were compared to a null, intercept only, model in order to identify a 'winner' (using the *withmain* setting in the `BayesFactor` package, R. D. Morey &

Table 6.1: Log Bayes factors for estimates of sensitivity (A_g)

	Model	$\log(B_{M,0})$	% error
1	$A_g \sim$ PT + TT + AG + TT:AG + ID	32.54	0.65
2	$A_g \sim$ PT + TT + AG + ID	32.35	0.78
3	$A_g \sim$ TT + AG + TT:AG + ID	32.08	0.75
4	$A_g \sim$ TT + AG + ID	31.94	0.40
5	$A_g \sim$ PT + TT + PT:TT + AG + TT:AG + ID	31.94	1.05
6	$A_g \sim$ PT + TT + PT:TT + AG + ID	31.72	3.11
7	$A_g \sim$ PT + TT + PT:TT + AG + PT:AG + TT:AG + PT:TT:AG + ID	31.16	2.32
8	$A_g \sim$ PT + TT + AG + PT:AG + TT:AG + ID	30.90	1.75
9	$A_g \sim$ PT + TT + AG + PT:AG + ID	30.71	1.80
10	$A_g \sim$ PT + TT + PT:TT + AG + PT:AG + TT:AG + ID	30.28	1.23
11	$A_g \sim$ PT + TT + PT:TT + AG + PT:AG + ID	30.02	1.04
12	$A_g \sim$ PT + TT + ID	28.81	0.42
13	$A_g \sim$ TT + ID	28.43	0.14
14	$A_g \sim$ PT + TT + PT:TT + ID	28.10	0.79
15	$A_g \sim$ AG + ID	2.89	0.41
16	$A_g \sim$ PT + AG + ID	2.76	0.66
17	$A_g \sim$ PT + AG + PT:AG + ID	1.05	1.96
18	$A_g \sim$ PT + ID	-0.16	0.42

Note: All Bayes factors are relative to a null model with a random participant effect only (model 0: $A_g \sim$ ID). AG = Age-Group, PT = Presentation Type, TT = Test Type, ID = participant ID, and ':' denotes an interaction effect

Rouder, 2015). Table 6.1 presents the results of this analysis.

As the table shows, the 'winning' model included main effects of presentation type (intrinsic, extrinsic), test type (colour, shape, binding), and age-group (younger, older) as well as the interaction between age and test type. Model 2 did not include this crucial interaction and comparing these two models reveals that the evidence in its favour is rather weak ($B_{1,2} = 1.21$). In an analysis of A_z the weight of evidence was equivalently weak, but against the interaction ($B_{1,3} = 1.66$). To follow this up, additional analyses were conducted on A_g comparing (1) feature probes (average of colour and shape) or (2) shape alone to binding probes. For both of these analyses the top model did not include the age × test type interaction, with the null favoured by over 4-to-1 in the feature versus binding analysis. In the shape versus binding analysis the evidence is far weaker, and does not favour either model ($B_{1,2} = 1.09$). Figures 6.2 and 6.3 clearly show that, regardless of presentation type, the effect of age was greatest for shape probes.

The key interaction of interest here is the three-way interaction between all experimental factors. This can be evaluated by comparing the model including this interaction against the model excluding it. Model 7 included the age-group × test

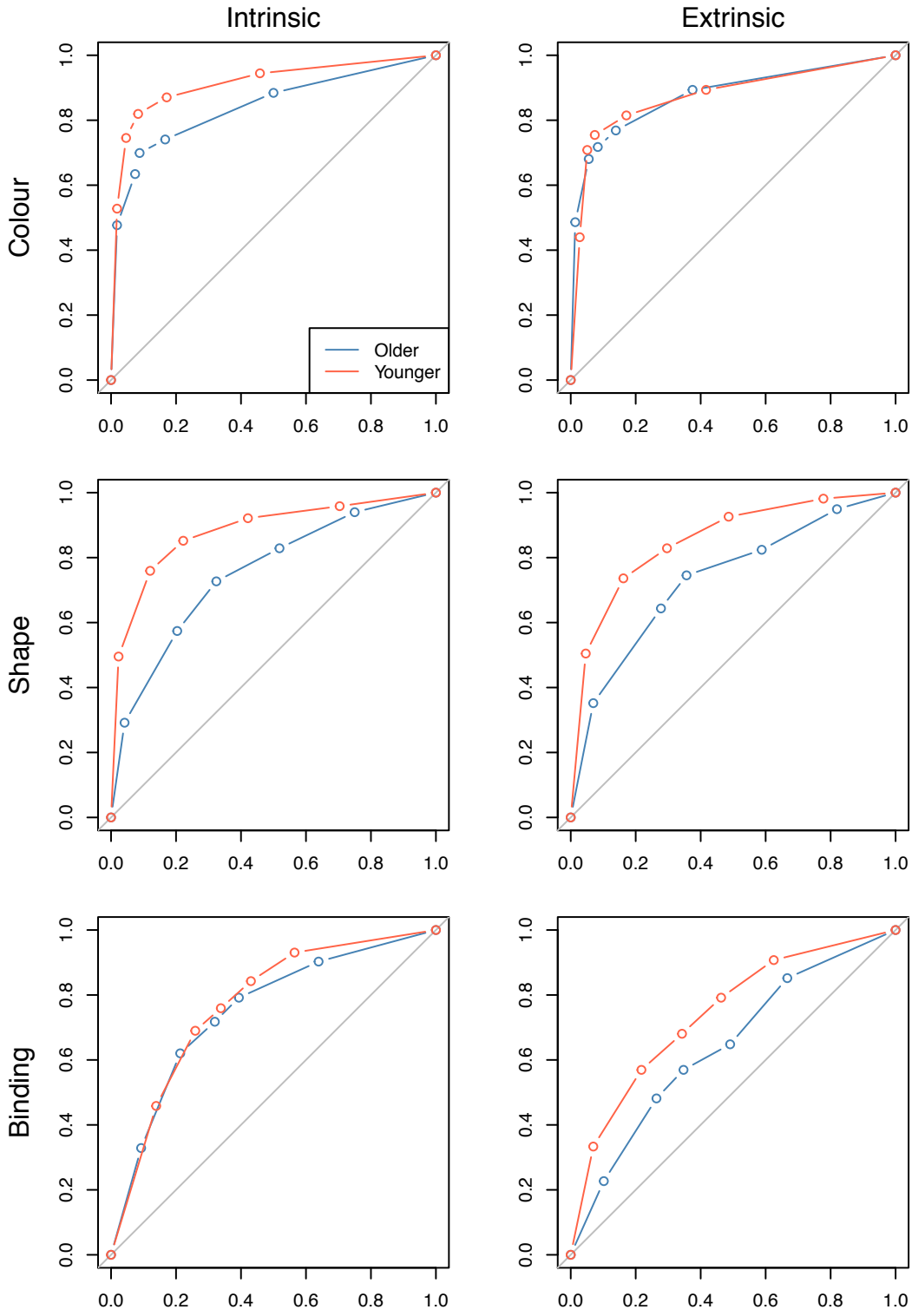

Figure 6.2: Empirical ROC curves across presentation and memory conditions for younger and older adults.

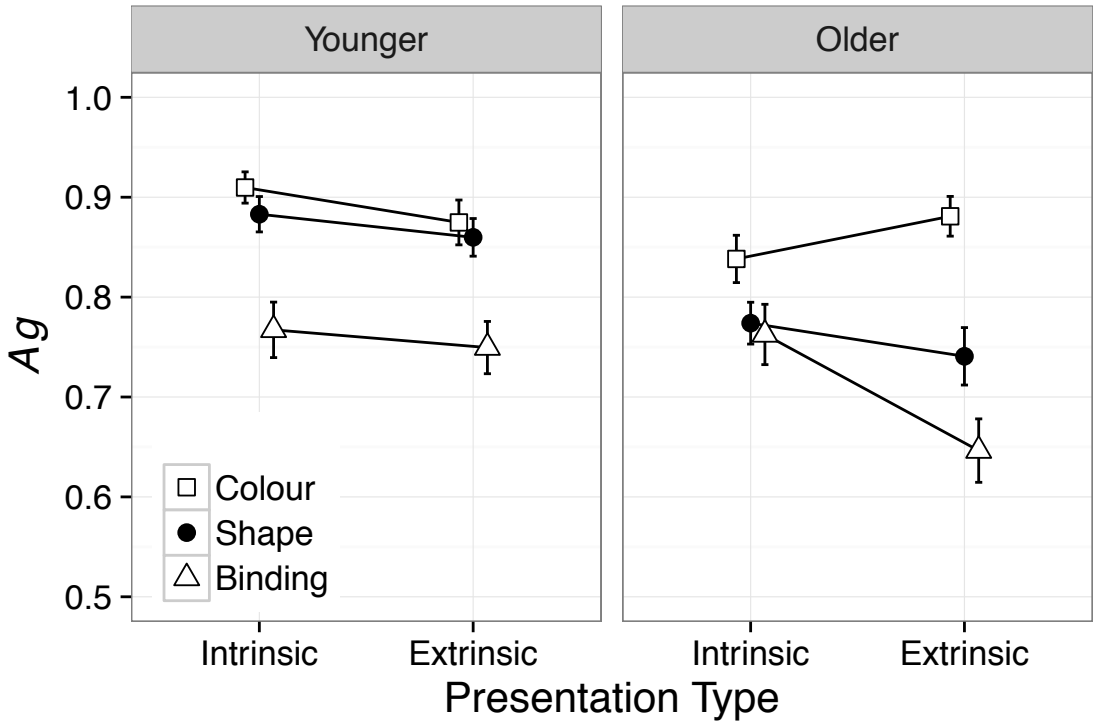

Figure 6.3: A_g across age groups and experimental conditions. Error bars display ± standard error.

type interaction whereas model 10 did not (see Table 6.1) and comparing these models revealed approximately 2-to-1 evidence in favour of the interaction. For A_z the weight of evidence was over 3-to-1. Building on this the separate, more focused, analyses revealed a similar level of support for this three-way interaction (feature versus binding = 2.6-to-1; shape only versus binding = 1.8-to-1). As can be seen in Figure 6.3 older and younger adults' sensitivity to binding changes was comparable in the intrinsic presentation condition whereas there was a clear difference in the extrinsic condition. This tendency was not present for the other test types. Thus we have very slight evidence that the presentation format modulated the size of the binding cost differently across the two age-groups. However, it is important to note that this is largely driven by poorer detection on *same* trials by older adults for extrinsic-bindings trials. This can be seen in the ROC curve (bottom right panel of Figure 6.2) which is shifted downwards (implying a lower hit rate). To follow this up we assessed the role of serial position in older adults' change detection accuracy.

Serial Position

As no-change trials involve the repetition of an item from the study sequence it is possible to look at the role of serial position in accuracy on *same* trials. Figure 6.4 plots accuracy across the presentation and test types for each item in the sequence. This experiment was not set up to assess the role of serial position and the data are rather sparse at this level (around 3 trials per position per condition). Consequently the data are not suited to a full model based analysis (with sparse data the posterior distribution is dominated by the prior). Thus the data are plotted with bootstrapped 95% confidence intervals (CIs) to enable 'inference by eye' (Cumming & Finch, 2005).

While there is fairly considerable uncertainty in the estimates of accuracy it appears that older adults' accuracy in the extrinsic binding condition (bottom-right panel of Figure 6.4) was particularly poor at serial positions 1 and 2. This is clearest at position 2 where the 95% CIs do not overlap and this tendency is not present in the younger adults' performance. On this basis we may tentatively suggest that older adults specifically struggled to hold extrinsic feature pairings in mind whilst processing incoming stimuli or they strategically chose to focus on the final item in the sequence and hold this in a privileged state (see, e.g., Hu, Hitch, Baddeley, Zhang, & Allen, 2014). Future work may extend the testing session to obtain a greater number of trials per serial position to look at this in more detail.

6.3 Discussion

There is overwhelming evidence for a specific age-related deficit in retaining associations in LTM (Old & Naveh-Benjamin, 2008a) and growing evidence that this is also the case for STM/ WM (T. Chen & Naveh-Benjamin, 2012; Hartman & Warren, 2005). This is in stark contrast to failures to demonstrate feature binding deficit in VWM (Brockmole et al., 2008; Isella et al., 2015; Read et al., 2016) and studies that find positive evidence *against* such deficits (Brown et al., 2016; S. Rhodes et al., 2016, see Chapters 3–5 in this thesis). This may reflect an important dis-

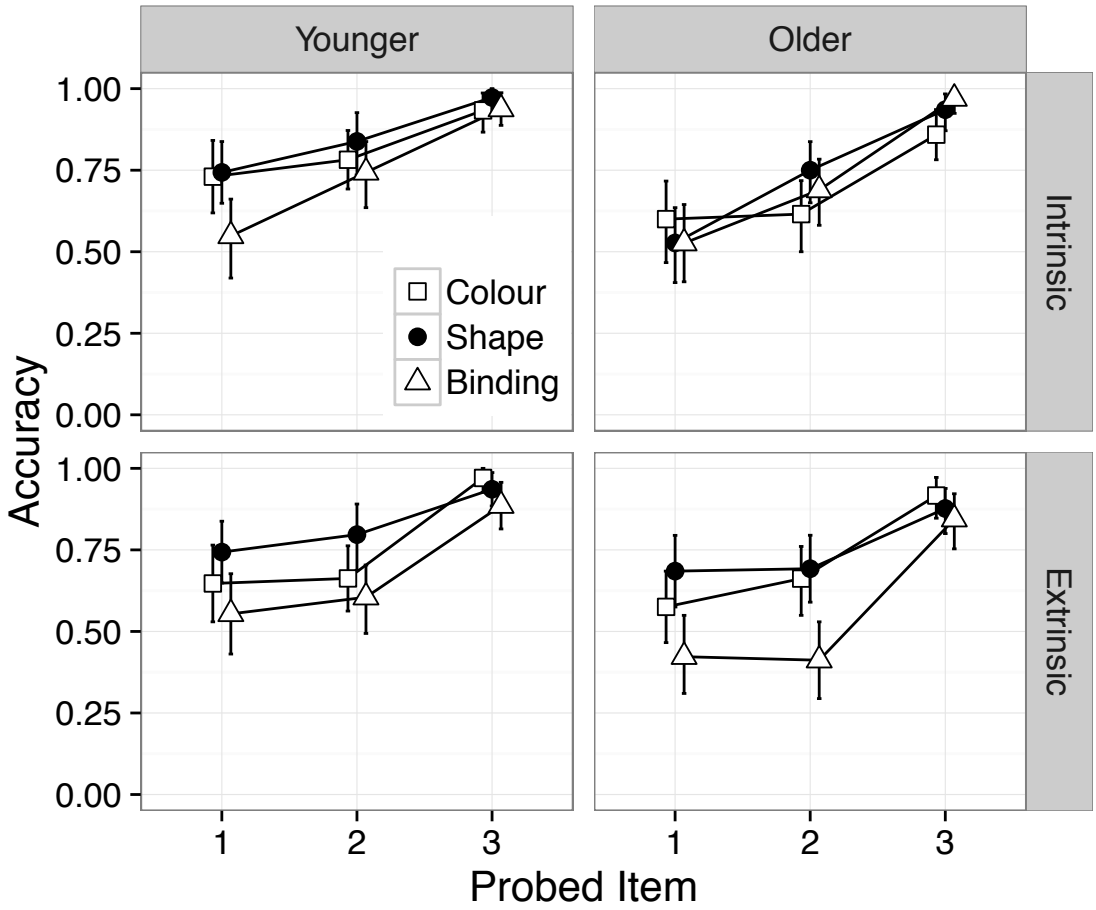

Figure 6.4: Accuracy on no-change trials depending on serial position probed. Error bars are 95% bootstrapped confidence intervals.

tinction between mechanisms responsible for binding features *intrinsic* to an object from those responsible for binding *extrinsic* contextual attributes (Zimmer & Ecker, 2010; Zimmer et al., 2006). Indeed, it seems the manner in which feature associations are presented greatly affects the ease with which they are remembered (e.g., Asch et al., 1960; Ceraso et al., 1998; Ecker et al., 2013). However, studies that do demonstrate age-related binding deficits also use very different stimuli to those that do not (T. Chen & Naveh-Benjamin, 2012) and, to date, little work has attempted to address this in an ageing context (although see, van Geldorp et al., 2015). Thus the present work attempted to contrast intrinsic and extrinsic binding mechanisms using identical stimuli in the same groups of younger and older adults.

Here we presented pairings of colour and shape either as integrated objects or as part of distinct items and obtained independent measurements of memory for

the individual features and their associations. Overall older adults appeared to be specifically poor at detecting changes to shape only, regardless of the mode of presentation (although the evidence for the interaction was weak). Increasing the size of the pool of stimuli from which the shapes were drawn may have increased the number of comparison errors made at test, particularly for the older adults (Noack et al., 2012; Pertzov et al., 2015). Further, older adults' sensitivity to binding changes was rather poor when the colour and shape were presented extrinsically (Figure 6.3). Bayes factors provided weak evidence in favour of the three-way interaction (relative to a model excluding this component). A preliminary examination of serial position curves (Figure 6.4) suggested that older adults experienced difficulty in retaining the first two extrinsic pairings in the study sequence, whereas the final pairing was well remembered. Thus this experiment gives some initial support to the notion that older adults struggle in relational binding and more specifically that these representations are potentially more fragile and susceptible to interference from subsequent memoranda. However the weight of evidence for the three-way interaction underlines the need for additional data.

It is important to note that the pattern of results here is somewhat out of step with the associative deficit found in LTM. As shown in the bottom right panel of Figure 6.2, older adults in the extrinsic binding condition were less likely to detect intact feature combinations (i.e. lower accuracy on *same* trials) whereas there appears to be less of an age-effect on the detection of feature swaps. The associative deficit, on the other hand, is characterised by an increased tendency towards false recognition of recombined lures, which has been attributed to over-reliance on familiarity with the component features (e.g., T. Chen & Naveh-Benjamin, 2012; M. G. Rhodes et al., 2008, see Chapter 1). Why extrinsic presentation in the present study had an effect on older adults' recognition of intact bindings is unclear. Accuracy across serial positions suggests that earlier feature pairings may have suffered interference from subsequent items or that older adults strategically focused on the final pairing. However, a larger number of trials will be necessary in future work to assess this. When binding deficits are observed in WM it is important to establish whether they

are qualitatively similar to those observed in LTM—close investigation may reveal different underlying causes.

While we found suggestive evidence that older adults struggle to retain extrinsic features in VWM, the earlier study of van Geldorp et al. (2015) did not. In their experiment the effect of age (young, middle, old) was no greater when recalling colour-shape combinations presented as spatially distinct items versus when they were presented as a conjunction—the evidence *against* the age-group by binding type interaction is not clear, however. In this study, unlike the present one, the intrinsic and extrinsic (conjunctive and relational) presentation formats were blocked. This may have allowed participants to adopt conjunctive encoding strategies in the relational condition (for example, imagining the colour filling the shape). These intentional strategies can greatly affect performance (Ecker et al., 2013) and the size of age-related binding deficits (Bastin et al., 2013). In the present study observers were unaware (at least for the first memory screen) of what presentation format to expect on a given trial and therefore may have been less able to adopt these potentially blurring strategies. It is interesting to note that in van Geldorp et al. (2015) the binding performance of younger adults appeared to correlate with performance on neuropsychological tests assessing MTL function, whereas older adults' recall performance correlated with frontal measures, possibly pointing towards the use of effortful strategies. The use of sequential presentation here may also have revealed a specific weakness of older adults' memory for extrinsic feature bindings. Figure 6.4 suggests a particular difficulty in recognising feature associations presented early in the sequence with relatively preserved retention of the final pairing. It is hoped that this study will provide the basis for future explorations of different binding mechanisms in healthy ageing.

As outlined in the Introduction, there is growing evidence that intrinsic and extrinsic binding are subserved by differing brain networks (e.g., Piekema et al., 2010; Parra et al., 2014). Namely extrinsic binding appears to depend on the MTL, specifically the hippocampus (Parra, Fabi, et al., 2015), whereas intrinsic binding is spared by hippocampal lesions and has been found to activate occipital/ parietal

areas (e.g., Parra et al., 2014; Xu, 2007). Thus our findings may be predicted on the basis of both cross-sectional and longitudinal assessment of brain volume suggesting a particular effect of healthy ageing on the hippocampi (see, Raz & Rodrigue, 2006, for a review).

Interestingly, there is evidence that this anatomical division is important for understanding the conjunctive binding deficit in early Alzheimer's disease (AD; Della Sala et al., 2012; Parra, Abrahams, Fabi, et al., 2009; Parra, Abrahams, Logie, Mendez, et al., 2010). In reviewing the literature, Didic et al. (2011) suggested that context free, object memory is one of the first memory systems to deteriorate in AD given pathological change to extra-hippocampal regions of the anterior MTL (perirhinal and entorhinal cortices). Bastin et al. (2014) recently looked at this by measuring resting cerebral metabolic rate with PET in a group of mild-AD patients and healthy controls. Using the same encoding manipulation as their earlier ageing study (Bastin et al., 2013, see above), they found that hypometabolism in the left parahippocampal gyrus and anterior extra-hippocampal regions of the MTL predicted patients' recall of colour encoded as part of an object (i.e. intrinsic). Whereas the deficit for recalling colour encoded as context was associated with regions of the default mode network (including anterior medial PFC and precuneus), commonly associated with episodic memory (see, e.g., Buckner et al., 2005). At a structural level, Das, Mancuso, Olson, Arnold, and Wolk (2015) recently found a posterior-anterior MTL divide in a group of participants with amnestic mild cognitive impairment in relation to WM performance. Grey matter volume in the anterior MTL was related to object change detection, whereas posterior MTL volume was related to object-context change detection. Finally, the integrity of inferior frontal white matter and corpus callosal tracts has been found to predict WM binding performance of individuals with familial AD on tasks similar to that used in the present work, whereas paired associates learning is predicted by the integrity of hippocampal white matter projections (Parra, Saarimäki, et al., 2015).

Thus the early pathological change to anterior MTL can be observed with neuroimaging techniques and may underlie the specific conjunctive binding deficit seen

in early AD (Didic et al., 2011; Parra, Abrahams, Logie, Mendez, et al., 2010). Future work assessing functional and structural changes associated with AD and their relation to binding deficits may also consider assessing measures of recollection and familiarity. As discussed above, a recent meta-analysis points to AD-related deficits in both recollection and familiarity, whereas healthy ageing seems to largely spare familiarity (Koen & Yonelinas, 2014). As these processes map theoretically onto extrinsic and intrinsic binding (Zimmer & Ecker, 2010) and have been associated with overlapping brain regions (Diana et al., 2007; Eichenbaum et al., 2007) pursuing this link may result in more efficient behavioural and physiological markers for AD.

In summary, building on previous work (van Geldorp et al., 2015), we assessed whether older adults specifically struggled to bind simple features (colour and shape) in VWM when these features were extrinsic to each other, that is came from separate objects, as opposed to being presented within the same item. Obtaining independent measures of item and associative memory and reducing the potential influence of strategy use when a specific presentation format is expected, we found suggestive evidence for this suggestion. Although it should be reiterated that the pattern of performance found here is not exactly that predicted by an associative deficit. Further work is needed to better characterise the nature of WM binding deficits and their relation to those observed in LTM. The method outlined here will provide a useful starting point for this future work.

Chapter 7

Age-Related Decline of VWM: Exploratory Modelling

7.1 Introduction

The present thesis has primarily focused on the issue of whether older adults exhibit a *specific* deficit when required to integrate and retain multi-featured stimuli in VWM. Across several experiments with almost 300 participants we have found no such evidence. Chapter 3 reported an experiment assessing a potential role for presentation time suggested by previous work (Brown & Brockmole, 2010) and found good evidence *against* an age-related binding deficit. Chapters 4 and 5 report two large experiments (by comparison to the rest of the literature) addressing the question of whether healthy older adults are more likely to miss conjunction changes when they are mixed in with more salient changes to component features. In contrast to previous work (Cowan et al., 2006) we found strong evidence against this being the case (see also, T. Chen & Naveh-Benjamin, 2012). In Chapter 5 we assessed binding between colour and location, as opposed to binding between the surface features of objects, and found, contrary to previous suggestions (Borg et al., 2011; Mitchell, Johnson, Raye, Mather, & D'Esposito, 2000; Mitchell, Johnson, Raye, & D'Esposito, 2000) but in line with other recent findings (Pertzov et al., 2015; Read et al., 2016), that older adults do not struggle to retain what was where

any more than they do maintaining what *or* where.

Beyond what we *did not* find, what we did consistently find was a large effect of age on change detection performance. This was observed both in terms of raw accuracy and in measures of sensitivity (e.g. P_r) and estimates of the number of items in VWM (k) in all conditions and at all set sizes used in the present work. This is certainly not new as there are now numerous reports on the age-related decline of visual working memory using the change detection task (e.g., Jost et al., 2011; Ko et al., 2014; Sander et al., 2011a, 2011b; Sander, Werkle-Bergner, & Lindenberger, 2012). This is also in line with the findings from more conventional span type measures which suggest that older adults are able to retain less information in working memory (Bopp & Verhaeghen, 2005; Verhaeghen, Marcoen, & Goossens, 1993), although the exact magnitude of the age-effect depends greatly on the type of material used (e.g. W. Johnson et al., 2010, see Chapter 1 for more detail). There have been many accounts of the age-related decline of VWM performance that tend to emphasise either a reduction in the overall size of VWM storage or less efficient management of VWM resources. More recently the importance of fluctuations of attention for change detection performance, and performance on WM tasks more generally, has been increasingly addressed by researchers (Adam, Mance, Fukuda, & Vogel, 2015; Unsworth & Robison, 2016) but it is not clear to what extent this factor contributes to age-differences in performance. However, at least one recent study has attempted to separate out the contribution of the number of items that can be effectively held in VWM and lapses of attention to age-differences in change detection performance (Sander et al., 2011a). The current Chapter aims to build on this and presents the results of exploratory modelling of the data from Chapters 4 and 5. First however, we give a summary of popular accounts of age-related WM decline and the growing interest in how lapses of attention contribute to individual differences in performance.

Accounts of Age-Related VWM Decline

Theoretical accounts of poorer performance on working memory tasks with advancing age are, unsurprisingly, numerous. However, they tend to suggest either a reduction in the amount of VWM capacity available for the storage of memoranda or less efficient management of VWM resources (or both). An example of the former is the suggestion that healthy ageing reduces the *precision* with which representations can be stored (e.g. Noack et al., 2012). In this case VWM capacity is conceptualised as a flexible resource with greater resource allocated to an item resulting in a greater signal-to-noise ratio and, consequently, a more precise representation. In line with this suggestion, older adults appear to require larger magnitude changes in order to accurately perform change detection tasks (Noack et al., 2012) and tend to exhibit greater variability in recalling features from VWM (Peich et al., 2013; Pertzov et al., 2015). Further, this account also has a viable biological origin in the senescent change of the dopaminergic neurotransmitter system (see Li & Sikstrom, 2002; Störmer, Passow, Biesenack, & Li, 2012, for reviews). The stimuli used in many VWM tasks are categorically distinct (or supra-threshold) and for younger adults appear to be sufficient to support an all-or-nothing discrimination process (Donkin et al., 2013; Rouder et al., 2008). Nevertheless, it is not clear if this is also the case for older adults. Thus lower capacity estimates in our change detection task may reflect difficulty in comparing the probe item to an internal representation distorted by noise.

Other accounts have implicated not a change in VWM capacity per se, but a reduction in efficiency with which VWM capacity is used. These accounts tend to refer to deficits in the 'top-down' control of VWM contents (Gazzaley, Cooney, Rissman, & D'Esposito, 2005; Gazzaley et al., 2008; Jost et al., 2011; Sander et al., 2011b; R. West, 1999) associated with deterioration of the frontal lobes (Raz & Rodrigue, 2006; R. L. West, 1996) (however see, S. Rhodes & Parra, 2016, for a critique of equating executive control with the frontal lobes). For example, Jost et al. (2011) used EEG recording in conjunction with a whole display change detection task, in which younger and older adults had to detect changes to orientation. However, for

some trials there were stimuli in the memory array that were task-irrelevant and had to be ignored. The authors assessed the contralateral delay activity (CDA)—a difference waveform obtained by subtracting the amplitude at occipital electrodes ipsilateral to memory stimuli from those contralateral, which had been previously shown to relate to behavioural measures of VWM capacity (Vogel & Machizawa, 2004; Vogel, McCollough, & Machizawa, 2005)—and found that older adults were less able to ignore irrelevant distractor stimuli particularly at early stages of the retention interval, as evidenced by greater amplitude of the CDA when distractors were present (however see Parra et al., under review; Ko et al., 2014, for a criticism of the CDA as a neural index of capacity). While older adults do appear to struggle to filter out task irrelevant information, in our present tasks all stimuli were relevant to the task therefore the demand to filter out irrelevant information at encoding was minimal.

Rather it may be the case that older adults are particularly susceptible to the effects of *proactive interference* (PI; Hasher & Zacks, 1988). It may be that our adults were less able to update the contents of VWM on each trial and some capacity was expended on memoranda from previous trials. While some have argued that PI does not play a particular role in change detection performance (Lin & Luck, 2012; Logie et al., 2009) other work, in particular that manipulating the temporal distinctiveness of trials, has detected the influence of interference between trials (Hartshorne, 2008; Makovski & Jiang, 2008; Shipstead & Engle, 2013). Thus an age-related decline in the ability to efficiently update the contents of WM (see, S. Rhodes & Parra, 2016; Verhaeghen, 2011, for reviews) would lead to poorer change detection performance given the presence of items in VWM from previous trials. Indeed there is evidence for an influence of PI on age differences in working memory capacity estimates. For example, age-differences in WM span are larger for sequences of trials that draw stimuli from the same semantic category (and therefore more likely to mutually interfere) as opposed to trials following a category change (Emery, Hale, & Myerson, 2008). Further, Bowles and Salthouse (2003) found that, when accounting for individual differences in susceptibility to PI, age

related differences in WM span were reduced by approximately half.

Of course, change detection performance, and performance on tasks assessing WM more generally, are likely to reflect the output of multiple interdependent processes (Logie, 2011) as well as different strategic approaches to the task at hand (Logie, Della Sala, Laiacona, Chalmers, & Wynn, 1996). Thus it seems likely that, rather than one of these factors alone, some combination of the above factors determines age-differences in the estimated capacity of VWM (for an example of such an account, see Sander, Lindenberger, & Werkle-Bergner, 2012).

Another important source of variation in VWM capacity estimates, that has received little attention to date, is that of lapses of attention (Adam et al., 2015; Unsworth & Robison, 2016). Here, rather than attributing failures of change detection to differences in the use of attention to manipulate and update VWM resources, focus shifts to possible age-differences in the ability to sustain attention across a series of trials. Recently, Unsworth and Robison (2016) found that younger participants reporting a greater frequency of task unrelated thoughts (TUTs) when probed tended to obtain smaller estimated VWM capacity from a standard change detection task. Even more interestingly, when combined with a change detection task in which participants had to ignore distractors (Experiment 3) the ability to filter out irrelevant information appeared to contribute *independent* variance to VWM capacity. Thus, based on these initial findings, it appears that the ability to sustain attention on-task is largely unrelated to the ability to filter information in predicting VWM performance. To what extent healthy ageing modulates the probability of lapses in attention is an emerging area of research that we attempt summarise below.

Age differences in lapses of attention

Relative to younger adults, older adults are less likely to report mind wandering or TUTs during tasks requiring sustained attention. For example, on the sustained attention to response task participants perform a go/ no-go type task with regular probes asking them about their thoughts (either related to the task or not) imme-

diately preceding a stimulus and older adults have been shown to produce fewer task unrelated thoughts (Carriere, Cheyne, Solman, & Smilek, 2010; McVay, Meier, Touron, & Kane, 2013; Jackson & Balota, 2012). This, perhaps counterintuitive, finding has been corroborated by both behavioural results showing fewer responses on no-go trials (Carriere et al., 2010, but slower overall responding) and recording of eye-movements during trials on which the mind wandered showing qualitative similarities to those of younger adults (Frank, Nara, Zavagnin, Touron, & Kane, 2015). Older adults also appear to experience more thoughts related to their performance on the task (i.e. self evaluation), however, it is unclear how these thoughts relate to actual task performance (Frank et al., 2015; Zavagnin, Borella, & De Beni, 2014). Overall, these findings may suggest minimal involvement of attentional lapses in the are-related decline of VWM.

However, it is worth noting that the tasks used in many studies of mind wandering use are usually fairly simple, for example reading a passage of text or performing a go/ no-go task, and age differences are typically not found on these tasks in terms of performance or older adults are slower to respond. Thus it may be the case that younger adults report TUTs with greater frequency as they are merely not being taxed enough by the current task. Indeed the suggestion that age-differences in mind wandering may have been overestimated has been made elsewhere (Maillet & Schacter, 2016; Zavagnin et al., 2014). Unfortunately age-differences in TUTs during more complicated tasks with multiple stages, such as the VWM change detection task, have not been assessed. However, it is interesting to note that Unsworth and Robison (2016) reported approximately 27% of their thought probes during their change detection were task unrelated; this contrasts with estimates typically greater than 40%—and as high as 70%—in younger adults with the typical go/ no-go tasks (e.g. McVay et al., 2013; Jackson & Balota, 2012).

Beyond work on mind wandering recent computational modelling of free recall has implicated age-differences in sustained attention as a key source of age-related decline in episodic memory (Healey & Kahana, 2016). Using an extension of the retrieved context framework, Healey and Kahana (2016) were able to account for

the complex pattern of age-differences in episodic free recall via four components; sustained attention across learning trials, the ability to form associations between content and context, retrieval monitoring, and noise associated with retrieval decisions. The fact that sustained attention was required, along with these other factors, in order to accurately reproduce the data pattern underlines a crucial role in age-related memory differences. Further, this four component model was also able to reproduce age-difference in a long-term memory recognition task.

Importantly, it is possible to separate out the contribution of attention lapses from that of VWM capacity to age-differences in change detection performance. Rouder et al. (2008) extended the model proposed by Cowan (2001) for the standard single probe change detection task to include the probability that the observer pays attention on a given trial. This additional parameter significantly improved model fit, accounting for the appearance of errors at low set sizes across a large number of trials. Applying such a model in the context of healthy ageing will indicate the extent to which a complicated explanation of performance differences (as outlined above) is needed relative to the simpler explanation that older adults are less able to sustain attention across the course of an experimental session.

Previous modelling of age-differences in change detection

To our knowledge only one study has used the measurement models described by Rouder and colleagues to separate out the contributions of attentional fluctuation, capacity, and guessing bias to age-differences in change detection performance.

Sander et al. (2011a) were interested in the relative contributions of low level binding processes and higher level strategic operations to change detection performance across the lifespan. In order to isolate these two components they assessed measures of capacity across three presentation times (100, 500, and 1000 ms) in the presence or absence of to-be-ignored distrator stimuli. In their change detection task participants studied between 2 and 10 to-be-remembered coloured squares and were probed using a whole display in which a single colour could have changed. Duplicates were allowed with the constraint that no colour could appear more than twice

and changes *could* occur to (or introduce) duplicates, thus requiring participants to bind each colour to its respective location. It is important to note here that Sander et al. (2011a) did not assess performance for component features and consequently it is not clear whether older adults' poorer change detection performance arose due to their VWM for the individual features or their combination. Their use of the term binding appears to refer more to the process of creating representations of items, regardless of the number of features, in VWM.

To separate out the contributions of capacity, lapses of attention, and guessing bias to task performance Sander et al. (2011a) applied the modelling approach introduced by Rouder et al. (2008). Estimates of VWM capacity were consistently lower in older adults relative to younger adults and remained so at longer presentation times (see Chapter 3). There were no obvious differences between younger and older participants in terms of the attention parameter suggesting that the probability of a lapse was approximately similar. Finally, younger adults were observed to exhibit a more conservative guessing bias. These results provide valuable insight into potential explanations of healthy older adults' poorer change detection—older adults may be less able to efficiently allocate VWM resources or have less space to fill whereas the ability to sustain attention on task is unaffected.

However, there are reasons to approach the findings of Sander et al. (2011a) with caution. As noted above the task used in the experiments of Sander et al. (2011a) was *whole display* change detection, but the model fit to the data was Rouder et al. (2008)'s extension of the processing model proposed by Cowan (2001). The appropriate model for the whole display would be a modification of the model proposed by Pashler (1988) (see R. D. Morey, 2011, for this extension). Rouder et al. (2011) point out that use of the incorrect, or unprincipled, model can greatly alter conclusions regarding the capacity of VWM and its covariates. Further, the use of duplicated colours is potentially problematic given that the assumption implicit in processing models of VWM is that items are categorically distinct and cannot be grouped into chunks. The presence of duplicates in the memory arrays of Sander et al. (2011a) may artificially inflate capacity estimates. With this in mind we

attempted to build upon the findings of Sander et al. (2011a) using data collected for Chapters 4 and 5, whilst taking advantage of recent developments in processing models of VWM (Cowan et al., 2013) to provide a more principled evaluation of the contribution of lapses of attention and capacity to age-differences in VWM.

7.2 Modelling Approach

The appropriate processing models for the single probe task used here are given in Chapter 2. As Rouder and colleagues (Rouder et al., 2008, 2011) point out, these basic models make the problematic prediction that performance at small set sizes (those within most individuals' capacity) should be flawless. However, it is clear that even at small array sizes, such as 2 or 3 items, observers can make errors due to lapses in concentration (suggesting that $k < N$). Rouder et al. (2008) modelled lapses of attention explicitly and we can similarly extend the models in Chapter 2 to accommodate an extra parameter. Extending the models in this way means that they must be fit to the data directly (either by using maximum likelihood estimation or MCMC methods) and we must make explicit assumptions about the way in which observers guess in this task. Rouder et al. (2011) distinguish between an observers' uninformed guessing rate, u, and guessing that is informed, g, by the observers' knowledge of k and the set size (N). They showed that it is possible that guessing is informed in the standard whole display paradigm but this could not occur for the standard single probe paradigm where items are probed in location. Here we show that in the single probe task, introduced by Wheeler and Treisman (2002), guessing *can be* informed and that this is qualitatively different depending on whether individuals are monitoring features or bindings.

Single Probe Informed Guessing

In the standard single probe change detection paradigm the probed item is either the same item (e.g. colour) studied *at that location* or is a brand new item. Assuming that observers use location to guide the discrimination (although see, Cowan et al.,

2013; Gilchrist & Cowan, 2014), when the probed item is outside VWM guessing cannot be informed. However, the single probe task used in the present work is different and allows for the possibility that guessing is informed by both k and N.

In the task introduced by Wheeler and Treisman (2002), observers identify sameness if the probe item is in VWM. However, if a match is not detected this could be due to one of two states of affairs; 1) the probe may be different (i.e. not in the study set) or, 2) the probe may be the same but was not encoded into VWM. With knowledge of the number of items (features or bindings) in VWM and the number of items presented it is possible for observers to modify their base expectation of a change to obtain the subjective probability that a change has occurred given that no-match was detected between VWM and the probe. Modifying Rouder et al. (2011, pp. 327):

$$Pr(\text{change} \mid \text{no-match}) = \frac{Pr(\text{no-match} \mid \text{change})Pr(\text{change})}{Pr(\text{no-match})},$$

gives the informed guessing rate, but the precise way in which it is derived depends on whether the single probe is testing VWM for features or for feature bindings.

In the individual feature condition the probability that no-match is detected given that a change has occurred, $Pr(\text{no-match} \mid \text{change})$, is necessarily 1. The probability that no-match is detected can be re-written as, $Pr(\text{no-match}) = Pr(\text{no-match} \mid \text{change})Pr(\text{change}) + Pr(\text{no-match} \mid \text{same})Pr(\text{same})$. The probability that no-match is detected on a same trial is the same as the probability that the probe item is outside VWM, $1-d$ where $d = \min(k/N, 1)$. Finally, the observer's base expectation of a change trial, u, determines the remaining components of this equation. Therefore, the informed guessing rate, given that no-match has been detected, for individual features is,

$$g_{\text{f}} = \frac{u}{u + (1-d)(1-u)}.$$

In the binding condition there is more information available to the observer. The probability that no-matching features have been detected given that a change has occurred is the same as the probability that *both* of the objects contributing features to the probe are outside VWM, $Pr(\text{no-match} \mid \text{change}) = 1 - c$. The probability

that, given the probe is the same, no-match is found with representations in VWM is, $Pr(\text{no-match} \mid \text{same}) = 1 - d$. Using this knowledge to modify the uninformed guessing rate we find that informed guessing in the single probe binding condition is given by,

$$g_{\text{b}} = \frac{(1-c)u}{(1-c)u + (1-d)(1-u)}.$$

With this in mind we can extend the models presented in Chapter 2.

For individual feature conditions an observer may miss a brand new feature probe either because the array exceeded capacity or they lapsed in attention and then incorrectly guessed. Therefore, the probability of a miss is now given by,

$$1 - h = a(1 - g_{\text{f}}) + (1 - a)(1 - u),$$

where a denotes the probability that attention is payed on a given trial. Consequently, $1 - a$ provides an estimate of the 'lapse rate' and is of primary interest here. When no-change has occurred a false-alarm may occur either because the relevant information was outside VWM or the observer lapsed:

$$f = a(1 - d)g_{\text{f}} + (1 - a)u.$$

In the binding condition the participant correctly identifies a change if either or both of the objects that donated features to the probe are in VWM or if they guess correctly. A hit in this case occurs with probability,

$$h = a[c + (1 - c)g_{\text{b}}] + (1 - a)u.$$

For a single probe that is an identical conjunction to one studied, provided an observer payed attention, they will identify this if the probe is in VWM. If the relevant information is not present or attention was not payed an error will arise if the observer guesses *same*. Therefore,

$$f = a(1 - d)g_{\text{b}} + (1 - a)u.$$

Notice that when attention is paid on a particular trial guessing can be informed, but when a lapse occurs guessing is, by definition, uninformed. With the appropriate formulae outlined we can turn to details of model estimation.

Hierarchical Modelling

Rouder et al. (2008) and Sander et al. (2011a) used standard optimisation routines to obtain maximum likelihood estimates of parameters for each participant. However, hierarchical Bayesian methods, which have been applied throughout the thesis, offer a more efficient use of information. We adopt a similar approach to R. D. Morey (2011), who details a hierarchical implementation of the Cowan (2001) and Pashler (1988) formulae which has proven useful in studies of VWM and dual-task interference (C. C. Morey, Morey, van der Reijden, & Holweg, 2013). There are three levels to the model:

1. Each trial is modelled as a Bernoulli trial with the probability of success determined by k, u, and a for that particular trial and the correct formula (defined above).

2. Parameter values are determined by linear models with fixed (grand mean, deflections from grand mean) and random (participant variability) components.

3. Prior distributions are placed on the components of the linear models.

As negative k values are nonsensical and both a and u are constrained to fall in the interval $[0, 1]$ the linear models at level 2 are placed on *transformations* of the parameters. In the case of a and u this is the simple logit transformation used elsewhere in the thesis. Thus the attention and guessing parameters for a given trial, i, are determined as follows,

$$\begin{aligned}\text{logit}(a_i) &= \mu^{(a)} + \mathbf{X}_i^{(a)}\boldsymbol{\xi} + s_{j[i]}^{(a)} \\ \text{logit}(u_i) &= \mu^{(u)} + \mathbf{X}_i^{(u)}\boldsymbol{\gamma} + s_{j[i]}^{(u)},\end{aligned}$$

where $\mu^{(a)}$ and $\mu^{(u)}$ represent grand mean parameter values and $\boldsymbol{\xi}$ and $\boldsymbol{\gamma}$ are deflections from those grand means associated with main- and interaction effects of experimental factors. These deflections from grand mean are constrained to sum-to-zero by the design matrices, $\mathbf{X}^{(a)}$ and $\mathbf{X}^{(u)}$, which contain effects coded variables for each trial. The final component of each linear model accounts for random individual variability in the rate of attention and guessing. Each are assumed to be

(independently) normally distributed with means of 0 and standard deviations, $\sigma^{(a)}$ and $\sigma^{(u)}$, estimated from the data.

For capacity, we used the mass-at-chance (MAC) transformation advocated by R. D. Morey (2011) (see also, R. D. Morey, Rouder, & Speckman, 2008; Rouder, Morey, Speckman, & Pratte, 2007), which is $k(\kappa) = \max(\kappa, 0)$, where a linear model is placed on κ which can fall in the interval $[-\infty, \infty]$ and negative values are mapped onto zero. Thus the capacity for a given trial, i, is determined as follows:

$$\kappa_i = \mu^{(\kappa)} + \mathbf{X}_i^{(\kappa)}\boldsymbol{\eta} + s_{j[i]}^{(\kappa)},$$

with $\mu^{(\kappa)}$ reflecting the grand mean capacity, $\boldsymbol{\eta}$ determining the deflections from this grand mean associated with the effects coded variables in $\mathbf{X}^{(\kappa)}$, and $s^{(\kappa)}$ reflecting random participant variability in k. Again, the random participant effect is assumed to be Gaussian with a mean of 0 and an estimated standard deviation, $\sigma^{(\kappa)}$.

As a cue to interpreting these models consider the situation where we are assessing the effect of a single factor with 3 levels on VWM capacity. The design matrix, $\mathbf{X}^{(\kappa)}$, in this case contains two effects coded variables in which level 3 forms the 'reference group' coded -1; in the first column level 1 is coded 1 and level 2 is coded 0; in the second column this is reversed. The vector $\boldsymbol{\eta}$ also contains two elements; the first reflecting the deflection from the grand mean associated with level 1 of the experimental factor (i.e. a main effect) and the second reflecting the deflection associated with level 2 (note: that the deviation associated with level 3 $= -(\eta_1 + \eta_2)$). For the sake of explanation, let $\mu^{(\kappa)} = 1$, $\eta_1 = 0.5$, and $\eta_2 = 0.2$. In this case the expected capacity (transformed) in a trial coming from level 1 of the design $= 1 + 1(0.5) + 0(0.2) = 1.5$, whereas a trial from level 3 has expected capacity $= 1 + -1(0.5) + -1(0.2) = 0.3$. Further, say participant 33 has a lower VWM capacity than average, for example $s_{33}^{(\kappa)} = -0.4$. For this individual expected κ for level 3 of the experimental factor is -0.1. The MAC transformation converts this meaningless estimate to a k of 0 and, as R. D. Morey (2011) notes, allows an experimental manipulation (e.g. a concurrent task) to use *all* of an individual's capacity.

At level 3 priors are placed on components of the linear models. For grand mean parameters and defections from this grand mean normal priors were used with a broad standard deviation of 10 for the scale of capacity and log-odds. The mean of these normal priors were set to zero for the deflection parameters (reflecting prior ambivalence to the direction of effects) and the grand mean of the guessing parameter ($\mu^{(u)}$). The prior for grand mean capacity ($\mu^{(\kappa)}$) was centered on 2.5 in line with previous studies estimating capacity from the single probe change detection task (Cowan et al., 2013) and for the attention parameter the prior grand mean ($\mu^{(a)}$) prior was centered at 3, reflecting our expectation that participants would pay attention on the vast majority of trials. To reiterate the size of the prior standard deviations mean that the prior mean placement will have little effect on the resulting parameter estimates. For participant effects priors are needed for the standard deviation of the random effect on a given parameter. We used the same gamma distribution for all SD parameters ($\sigma^{(\kappa)}$, $\sigma^{(a)}$, and $\sigma^{(u)}$) with shape = 1.01005 and rate = 0.1005012. As described in Chapter 2 this is a vague, non-committal, prior distribution with a mode of 0.1 and standard deviation of 10 (Kruschke, 2015).

As with our logistic model, described in Chapter 2, we used JAGS (Plummer et al., 2003) to sample 50000 times from the joint posterior distribution of the model parameters following a burn-in period of 5000 samples (model code is given in Appendix C). Convergence on a stable distribution was established across 4 chains using the multivariate BGR statistic (Brooks & Gelman, 1998).

7.3 Results

The hierarchical model was fit to the blocked data for both the colour-shape (Experiment 6) and colour-location (Experiment 7) experiments separately, thus allowing us to contrast parameter estimates from two data sets with different overall levels of performance. For both of these data sets we assessed main effects of age on capacity (k), attention (a), and guessing rate (u). In addition for capacity we also assessed the main effect of memory condition and the interaction between age-group and memory condition. We see no reason *a priori* to assess these additional compo-

Table 7.1: Posterior quantities from hierarchical multinomial processing model for colour-shape data

			95% HDI		
Parameter	Mean	Median	lower	upper	ESS
$\mu^{(\kappa)}$	2.507	2.509	2.284	2.721	1971.567
η_1: (1) Shape	-0.245	-0.244	-0.399	-0.093	6309.444
η_2: (2) Binding	-1.339	-1.339	-1.494	-1.182	3233.051
η_3: (3) Older Group	-0.222	-0.220	-0.439	-0.008	1864.059
η_4: 1×3	0.162	0.164	0.010	0.313	6219.095
η_5: 2×3	0.056	0.057	-0.098	0.203	3570.540
$\sigma^{(\kappa)}$	0.390	0.379	0.202	0.614	1647.955
$\mu^{(a)}$	1.194	1.186	0.772	1.624	2106.140
ξ_1: Older Group	-0.690	-0.691	-1.127	-0.254	1963.066
$\sigma^{(a)}$	1.141	1.125	0.764	1.544	3799.690
$\mu^{(u)}$	0.086	0.086	-0.038	0.217	3964.498
γ_1: Older Group	-0.107	-0.106	-0.232	0.025	3818.559
$\sigma^{(u)}$	0.382	0.378	0.280	0.490	10148.029

Note: The effects coded variables were as follows: (1) Shape = 1, Binding = 0, Colour = -1, (2) Shape = 0, Binding = 1, Colour = -1, (3) Younger = -1, Older = 1. Interaction contrasts were products of these effects coded variables.

nents for attention and guessing especially given the inferential cost associated with increasing model flexibility.

Colour-Shape Data (Experiment 6)

For the colour-shape data set, a summary of the posterior distribution is given in Table 7.1. It is clear from this table that the main effect of age was credibly non-zero for both capacity and attention parameters. For capacity the overall estimated age difference was -0.445 [-0.879, -0.016] items and there was no evidence that this reduced capacity was more pronounced for features relative to bindings (-0.168 [-0.610, 0.293]). For the attention parameter the age difference on the log odds scale was -1.381 [-2.253, -0.507] with a wide HDI that clearly excludes zero. For the guessing parameter this was not the case a the HDI *did not* exclude zero (-0.213 [-0.464, 0.049]), although younger adults tended to exhibit a bias towards guessing *different*.

The parameters are presented in Figure 7.1 on their natural scale (i.e. a and u as

probabilities). In order to gauge the relative importance of these three contributory factors to age-differences in change detection performance standardised mean differences were calculated. For each step in the MCMC chain the difference between young and old on the parameter of interest was divided by the estimated standard deviation of the random participant effect for that parameter. The resulting standardised mean difference summarises the magnitude of the age-difference in terms of expected deviations due to individual differences for a given parameter—placing them on a comparable scale.

For both the capacity and attention parameters the standardised mean difference between age-groups was of a roughly similar magnitude, with a difference of -1.215 [-2.563, -0.008] for k and -1.236 [-2.050, -0.437] for a. However, the width of the HDIs surrounding these contrasts reveal important differences in the uncertainty of these estimates. For for the capacity mean difference the credible values span a wide range from negligible values of essentially no difference to differences of over 2 standard deviations. For attention the HDI range appears to rule out 'small' age differences and imply a medium-to-large effect size. More light is shed on this by the second data set.

For the guessing parameter there is no clear difference between younger and older adults (-0.568 [-1.250, 0.118]), although this data set cannot clearly rule out a large effect size. As shown in Figure 7.1 older adults were almost perfectly neutral, whereas younger adults exhibited a more liberal guessing strategy. This goes against the observations of (Sander et al., 2011a) and we return to this issue in the Discussion.

Colour-Location Data (Experiment 7)

Table 7.2 presents a posterior summary of the parameters estimated from the colour-location data discussed in Chapter 5. Again older adults' VWM capacity was smaller than that of younger adults (-0.769 [-1.172, -0.359]) and this did not differ depending on whether the change detection task probed recognition of features or feature bindings (-0.332 [-0.800, 0.148])—if anything the binding cost was smaller in the

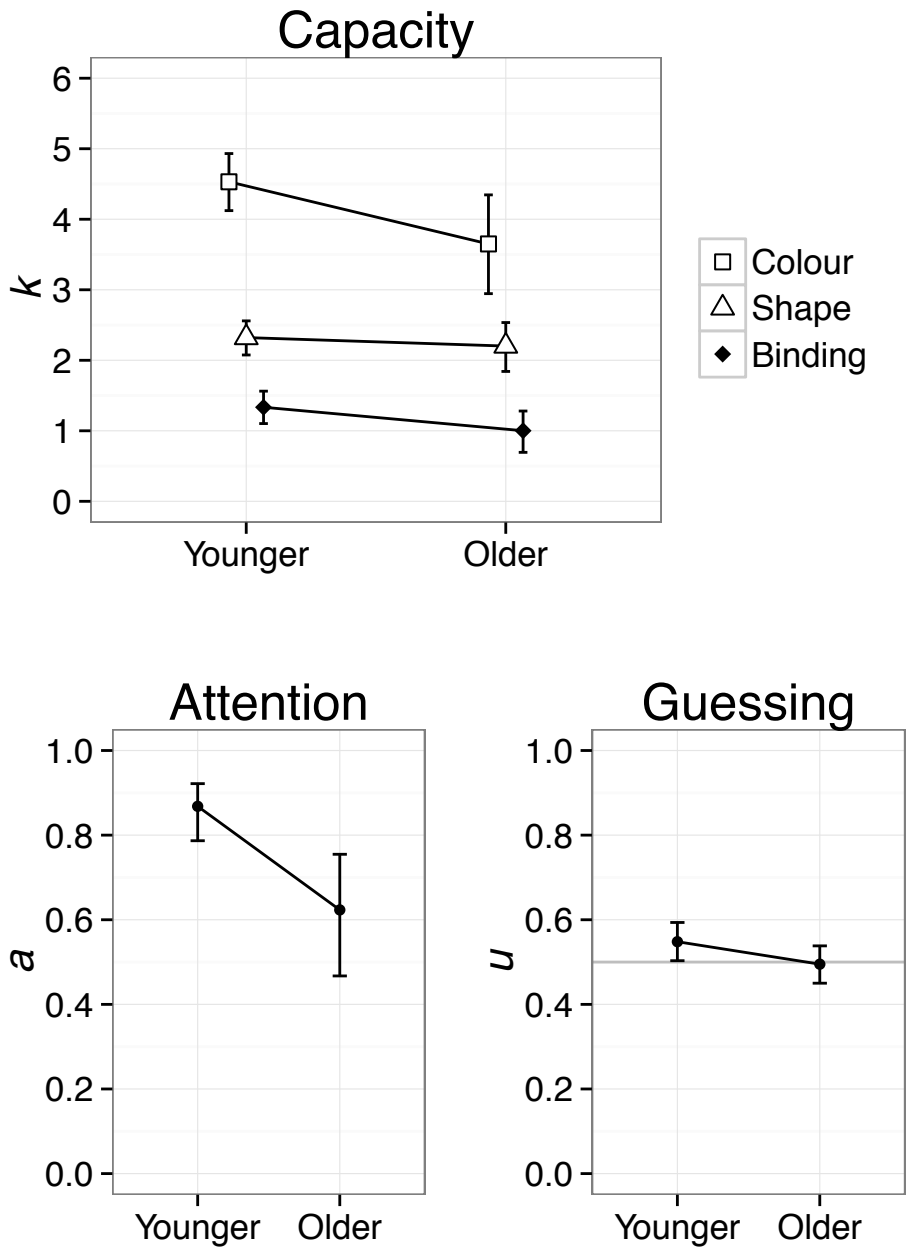

Figure 7.1: Results of exploratory modelling of the colour-shape data from Chapter 5. Points are posterior means and error bars denote the 95% Highest Density Intervals.

older group (see Figure 7.2).

There was a clear age-difference in the attention parameter (-0.656 [-1.176, -0.114]), however, relative to the analysis of the colour-shape data set, this difference was much less pronounced. Once again older adults appeared to adopt a somewhat more neutral guessing strategy but there was no clear evidence for an overall age-

Table 7.2: Posterior quantities from hierarchical multinomial processing model for colour-location data

			95% HDI		
Parameter	Mean	Median	lower	upper	ESS
$\mu^{(\kappa)}$	3.647	3.647	3.447	3.849	3350.912
η_1: (1) Location	1.169	1.169	0.989	1.345	10721.445
η_2: (2) Binding	-1.607	-1.608	-1.766	-1.442	5645.250
η_3: (3) Older Group	-0.384	-0.384	-0.586	-0.179	3369.180
η_4: 1×3	-0.115	-0.113	-0.298	0.062	11075.708
η_5: 2×3	0.111	0.111	-0.049	0.267	6795.007
$\sigma^{(\kappa)}$	0.473	0.467	0.297	0.661	3353.673
$\mu^{(a)}$	1.798	1.796	1.527	2.069	3287.583
ξ_1: Older Group	-0.328	-0.328	-0.588	-0.057	3320.131
$\sigma^{(a)}$	0.743	0.735	0.511	0.977	5635.771
$\mu^{(u)}$	0.399	0.398	0.229	0.571	4258.991
γ_1: Older Group	-0.121	-0.121	-0.291	0.046	4360.010
$\sigma^{(u)}$	0.493	0.488	0.355	0.640	8708.417

Note: The effects coded variables were as follows: (1) Location = 1, Binding = 0, Colour = -1, (2) Location = 0, Binding = 1, Colour = -1, (3) Younger = -1, Older = 1. Interaction contrasts were products of these effects coded variables.

effect on guessing bias (-0.242 [-0.581, 0.091]). Figure 7.2 depicts these trends.

In order to directly compare the magnitude of these age differences they were again standardised to a scale of expected individual differences. Although the raw mean difference in the attention parameter was smaller in the present data set relative to the colour-shape analysis above, the standardised mean difference was comparable in size (-0.900 [-1.641, -0.153]) implying a medium-to-large age effect on the probability of a lapse in attention. This data set also simultaneously supported a large effect of age on the capacity of VWM (-1.692 [-2.847, -0.638]) with the posterior mean implying an age effect above 1 standard deviation. Standardised age differences in guessing bias, as noted above, could not be clearly differentiated from zero (-0.500 [-1.198, 0.195]). Once again the HDIs accompanying these contrasts betray a large amount of uncertainty regarding the exact magnitude of these age effects but taken together this exploratory modelling points towards a role for *both* lapses of attention and reduced overall WM capacity in older adults' poorer change detection performance. The role of guessing bias is less clear but, if anything, older

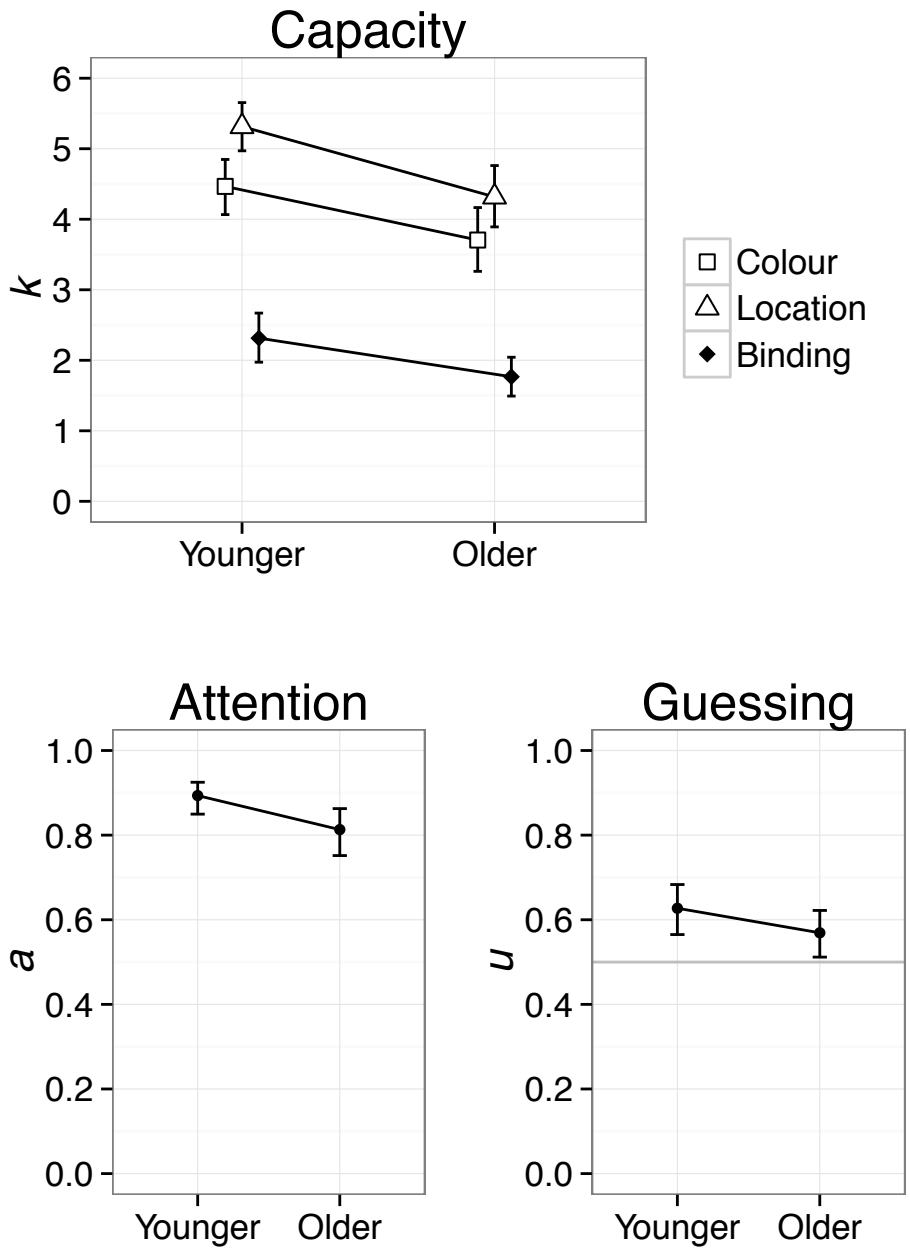

Figure 7.2: Results of exploratory modelling of the colour-location data from Chapter 6. Points are posterior means and error bars denote the 95% Highest Density Intervals.

adults appear to adopt a more efficient neutral position (see Figures 7.1 and 7.2).

7.4 Discussion

The present work has built upon a growing literature showing no additional age-related deficit in short-term/ working memory tasks requiring the maintenance of feature bindings. This work has also demonstrated a clear effect of age on VWM performance in line with a growing literature (e.g. W. Johnson et al., 2010; Jost et al., 2011; Ko et al., 2014; Sander et al., 2011a). As outlined in the Introduction to this chapter, there are many accounts of working memory decline that tend to explain age-differences in terms of reduced in WM resources (e.g. Noack et al., 2012) and/ or less efficient use of that capacity due to reduced executive, or top-down control (e.g. Sander et al., 2011b). However, the ability to sustain concentration and avoid lapses of attention is also an important contributory factor to performance on VWM tasks (Adam et al., 2015; Rouder et al., 2008; Unsworth & Robison, 2016). Whilst research on mind-wandering has suggested an age-related *decrease* in intrusive thoughts during tasks requiring sustained attention (e.g., Frank et al., 2015), recent computational modelling suggests a key role for sustained attention in age-related episodic memory decline (Healey & Kahana, 2016).

To our knowledge only one study has simultaneously assessed the role of lapses and capacity in age differences in change detection performance (Sander et al., 2011a). Applying a mathematical model (Rouder et al., 2008), Sander et al. (2011a) found an age-related decline in the number of items that could be retained in VWM but no clear evidence of an increase in lapses of attention. With the data collected for Experiments 6 and 7 (Chapters 4 and 5) we attempted to build on this previous work using a hierarchical version of the slots VWM model.

Across two data sets (one assessing colour-shape, the other colour-location) we found evidence for age-related decline in the number of items that could be retained in VWM and used to perform the change detection discrimination. *In addition* we found an increased lapse rate in our groups of older adults. Expressing these age-differences in terms of differences expected due to differences between individuals (i.e. standard deviations for each parameter) we find that in the colour-shape experiment standardised mean differences in both k and a were roughly comparable,

corresponding to an effect size of approximately 1.2 (-1.215 [-2.563, -0.008] for k and -1.236 [-2.050, -0.437] for a). However, the colour-shape data revealed a great deal of uncertainty surrounding the magnitude of these effects (as seen in the width of the HDIs). The colour-location data supported more firm conclusions and suggested that the effect of age on k (-1.692 [-2.847, -0.638]) was slightly greater than the effect on a (-0.900 [-1.641, -0.153]). Finally, although younger adults were generally more liberal when guessing, there were no clear age-differences in the probability of guessing 'different' when in an uncertain state.

Our findings regarding the number of items in VWM suggest that, even when differences in sustained attention are taken into account, older adults, on average, have the use of less information in WM. Unfortunately, our data are unable (and were not intended) to adjudicate between alternative explanations of reduced capacity estimates with age—however, we may speculate. It seems unlikely that this can be accounted for by interference from task unrelated stimuli, as all items in the memory array were relevant for the task. That being said, it is possible that, given task irrelevant features were present in both the memory and test arrays in individual feature conditions, older adults were more likely to encode these features, thus expending capacity on features unnecessary for task performance. This, however, would clearly predict a larger effect of age in the blocked conditions of the experiments reported in Chapters 4 and 5 relative to the condition in which trials were mixed (where features were always task relevant). The analyses reported in those Chapters provide good evidence against this, so we can rule this possibility out. Alternative explanations in terms of less precise representations (e.g. Noack et al., 2012) or a reduced ability to update the contents of WM leading to proactive interference (Bowles & Salthouse, 2003; Emery et al., 2008) are certainly viable and should form the basis for future work. These factors are not mutually exclusive, therefore this future work should aim to assess both at once—for example by varying both the temporal distinctiveness of trials and the magnitude of change—along with other potentially important factors that may modulate capacity estimates, such as age-differences in the use of strategy in WM tasks (W. Johnson et al., 2010; Logie

et al., 1996).

What contributed to our increased lapse rate? Anecdotal evidence from discussion with participants suggests that older adults were particularly influenced by the correct/ incorrect feedback given on each trial. It is possible that older adults, knowing that they had made a mistake on the previous trial, engaged more in post-error monitoring, perhaps increasing the likelihood of a lapse. This would certainly be in line with the finding that older adults report more task related interfering thoughts (McVay et al., 2013; Jackson & Balota, 2012; Zavagnin et al., 2014). Independent validation of this is needed, but a preliminary analysis of the combined raw data from Experiments 6 and 7 using `lme4` (Bates et al., 2014) predicting accuracy from previous trial accuracy (i.e. discarding the first trial of each block) and age-group revealed that, while previous trial accuracy predicted accuracy ($\beta = 0.135$, 95% CI [0.032, 0.238]), these two factors did not clearly interact (-0.028, [-0.193, 0.134])[1]. Thus, while there is some indication that errors were more likely to follow errors, there is no clear evidence in the present data that this tendency was stronger for older adults. Of course older adults produced more errors so it is still possible that post error monitoring could account for much of their increased lapse rate.

Alternatively, it is interesting to note that older adults were more likely to spontaneously report having mistakenly pressed the wrong response key. This might suggest that on a substantial portion of 'lapse' trials older adults had the relevant items in VWM but responded incorrectly due to temporary inattention at test. Of course it may be that younger adults made this kind of mistake as frequently but were less likely to spontaneously report this to the researcher. Future work explicitly requiring participants to report such errors or allowing participants to go back and correct mistakes may shed light on the origin of older adults' increased lapse rate.

Our findings diverge from those of Sander et al. (2011a) who only found age-difference in terms of the number of items in WM and no clear difference in the

[1]While accuracy was emphasised over speed in all of our experiments we also assessed evidence of post error monitoring in response time (RT). A mixed effects model was fitted to trials eliciting a RT less than 3 seconds, resulting in the loss of only 2% of trials. Previous trial accuracy predicted quicker RTs (-0.068, [-0.091, -0.045]) but this did not interact with age-group (0.011, [-0.024, 0.046]).

probability of lapsing. As noted in the Introduction the model applied by Sander and colleagues was unprincipled for their whole display task and is consequently highly likely to have given misleading parameter estimates (Rouder et al., 2011). The models used here, according to the slots model of VWM, are principled for the current single probe task (Cowan et al., 2013; H. Zhang et al., 2010) and accordingly give a better indication of the parameters and their covariates. With this in mind, the results reported here give a better indication of the factors underlying older adults' poorer change detection performance. Further our findings are in line with recent modelling of age-difference in episodic free recall and recognition which suggested multiple factors underlying older adults' poorer performance, including sustained attention (Healey & Kahana, 2016).

Limitations

One potential criticism of the present modelling is that we assumed that observers, in certain situations, use both the number of items in VWM and the set size to inform their guessing. There is some evidence that observers engage in probability matching in change detection tasks (Cowan et al., 2016; Rouder et al., 2008) however, to our knowledge, the crucial manipulations required to establish whether guessing is truly *informed* have not been performed (this would involve orthogonal manipulation of bias and set size across multiple levels). As a first approximation to assessing whether an uninformed or informed guessing model gives a better account of the data we also estimated a model in which all guessing was determined by the u parameter and compared the fit of the two models via the deviance information criterion (DIC). The DIC is commonly used to compare hierarchical models, where the number of parameters (and hence the flexibility of the model) is hard to directly quantify (Spiegelhalter, Best, Carlin, & Van Der Linde, 2002). JAGS uses MCMC sampling to estimate an 'effective number of parameters' (as described by Plummer, 2002, 2015) and adds this to the expected deviance of the model predictions from the observed data. For both the colour-shape and colour-location data sets the informed guessing model provided better fit to the data as evidenced by smaller DIC values

(ΔDIC = 100 and 247, respectively). As mentioned above, the crucial experimental manipulations needed to further elucidate the guessing strategy adopted by observers on difference versions of the change detection task are yet to be done but the present data set appears to be more consistent with an informed guessing model.

Another potential objection is that in the current case (like much of the previous modelling of VWM Donkin et al., 2013, 2014; R. D. Morey, 2011; Rouder et al., 2008; Sims, Jacobs, & Knill, 2012; van den Berg, Awh, & Ma, 2014) attention was modelled in an all-or-none fashion, where a lapse resulted in no information on the display items. However, fluctuations of attention may be more graded than this and produce variability in resulting capacity estimates. Indeed Adam et al. (2015) recently tracked whole report performance and showed that variability in the number of colours correctly reported was better accounted for by a graded attention model (see also, Cowan et al., 2016, for evidence of variability in k). Thus future work may adopt paradigms, like the whole report task, that give additional information that could be used to model age differences in the volatility of sustained attention.

Despite the limitations the present exploratory modelling is an improvement on previous attempts and provides an important starting point in refining accounts of age-related WM decline. Age-differences in the probability of inattention appear to make an important contribution to performance. Simultaneously older adults have access to less information from WM, which may result from reduced storage capacity, inefficient use of this capacity, or some combination of these factors.

Chapter 8

Group × Condition Interactions: Choice of Measure and Type I Errors

In experiments assessing recognition memory participants typically study a set of items (e.g. words or coloured shapes) and following a delay are required to distinguish previously seen items as *old* from previously unseen items which are *new*. The change detection task is no exception as participants must identify whether the probe is the *same* (i.e. *old*) or *different* (i.e. *new*). Performance on such tasks is captured by the frequency of *old* responses conditional on whether the probe was old or new. Participants make a *hit* if they correctly identify an old item, whereas they make a *false-alarm* if they incorrectly identify a new item as old. The frequency of hits and false-alarms will not only be influenced by an observer's ability to distinguish old and new items but also by their preference for one response option over another. Thus researchers regularly adopt measures that purport to separate out the sensitivity of an observer from their response bias. In order to do this, these commonly used measures make particular assumptions about the recognition process.

Two broad classes of model can easily be identified; 1) signal detection theory accounts propose that during a recognition task, items are evaluated on a continuous

decision variable (e.g. familiarity) and observers must establish a criterion, above which point they respond *old*, whereas 2) threshold accounts propose a small number of discrete states, where observers either detect a particular state of affairs or they are left in a situation of complete information loss and must guess whether an item has been encountered before.

What we aim to do in the current Chapter is to show that the choice between these alternatives is not arbitrary. Previous work has shown that the choice of measure, if not justified, can lead to erroneous conclusions regarding differences in sensitivity between two conditions (Rotello et al., 2008; Schooler & Shiffrin, 2005). The present work aims to add to this by considering tests of interaction effects, specifically Group × Condition interactions which have been the focus of the present thesis. This work is motivated by the fact that, while the experiments reported in Chapters 3–7 provide evidence against an age-related feature binding deficit, there have been previous such reports in the literature (e.g. Brown & Brockmole, 2010; Brown et al., 2016, Experiment 2; Peterson & Naveh-Benjamin, 2016; Isella et al., 2015). Despite this, the findings presented here have implications far beyond the feature binding literature to work on group differences in detection and recognition more generally.

We consider popular measures arising from simple incarnations of the SDT (d') and threshold (P_r, proportion correct) accounts. Of course there have been attempts to derive measures that eschew assumptions regarding the underlying recognition process and we consider a particularly popular measure in this vein (A'). Before describing the motivation for the present simulation and the way in which it was carried out the theory underlying these measures is introduced.

8.1 Signal Detection Theory

Accounts based on Signal Detection Theory (SDT) propose a graded decision variable (e.g. item familiarity) and that representations are distorted by noise; nevertheless old items tend to produce higher values on the decision variable resulting in two distributions that are separated to the extent that the observer can distin-

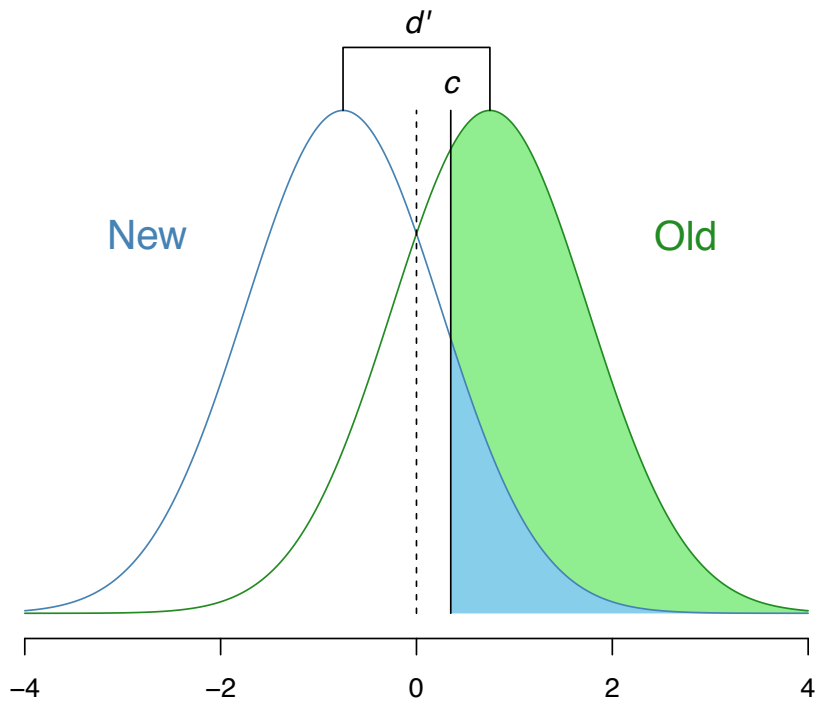

Figure 8.1: Illustration of Gaussian equal variance signal detection theory (GEV-SDT).

guish old from new items (Green & Swets, 1966; Swets, 1986b; Tanner Jr & Swets, 1954). As the observer has no direct access to the old and new evidence distributions, when an item is presented the elicited value is compared to some criterion; if the sampled value is above criterion the decision is *old* and if below, the decision is *new*. In the most commonly applied version of the SDT model the underlying distributions of the old and new items are assumed to be Gaussian and have equal variance (GEV-SDT). Figure 8.1 depicts this model.

In GEV-SDT d' quantifies the separation of the new and old evidence distributions in terms of their common standard deviation and thus the observer's sensitivity. The parameter, c, describes the criterion location relative to the intersection of the two distributions (dashed line in Figure 8.1) with negative values denoting a liberal response bias and positive values a conservative one. The probability of making a false-alarm is the probability that a value sampled from the new distribution falls above the criterion; for the GEV-SDT model this is given by, $f = \Phi(-\frac{1}{2}d' - c)$, where

Φ is the cumulative normal distribution function, and is shown by the blue shaded area of Figure 8.1. The probability of making a hit is shown by both the green and blue shaded areas and is given by, $h = \Phi(\frac{1}{2}d' - c)$. Combining and rearranging these terms gives an estimate of sensitivity,

$$d' = z(h) - z(f), \tag{8.1}$$

where z is the quantile function of the normal distribution ($z(x) = \Phi^{-1}(x)$). Similarly for criterion,

$$c = -\frac{1}{2}[z(h) + z(f)]. \tag{8.2}$$

Provided the assumptions of GEV-SDT are met Equation 8.1 provides a valid estimate of sensitivity uncontaminated by criterion placement.

A good way of visualising the predictions of different recognition models for our present purposes—one that has been in use since the inception of SDT (Tanner Jr & Swets, 1954)—is to plot the predicted hit rate as a function of false-alarm rate as response bias varies but sensitivity is held constant. The resulting plot displays the *isosensitivity*, or receiver operating characteristic (ROC), curve for a given measure. Figure 8.2A depicts ROC curves for the GEV-SDT model at various levels of sensitivity. The ROC is clearly non linear and symmetrical around the negative diagonal and, when plotted in z space, yields a linear function with a slope of 1 and intercept, d' (Macmillan & Creelman, 2005; Swets, 1986b).

While GEV-SDT is the most widely applied, it is certainly not the only account derived from SDT. Different distributional assumptions are possible, such as the logistic distributions underlying Luce's choice theory (Luce, 1963a). Further it is often observed that empirical ROCs from recognition experiments are asymmetrical and plotted in z space have a slope less than 1 (e.g. Ratcliff, Sheu, & Gronlund, 1992; Swets, 1986a). This can be accounted for by relaxing the assumption of shared variance between the old and new distributions and assuming that old items tend to produce more variability ($\sigma_o > \sigma_n$) as well as higher values on the decision variable (Macmillan & Creelman, 2005; Swets, 1986b).

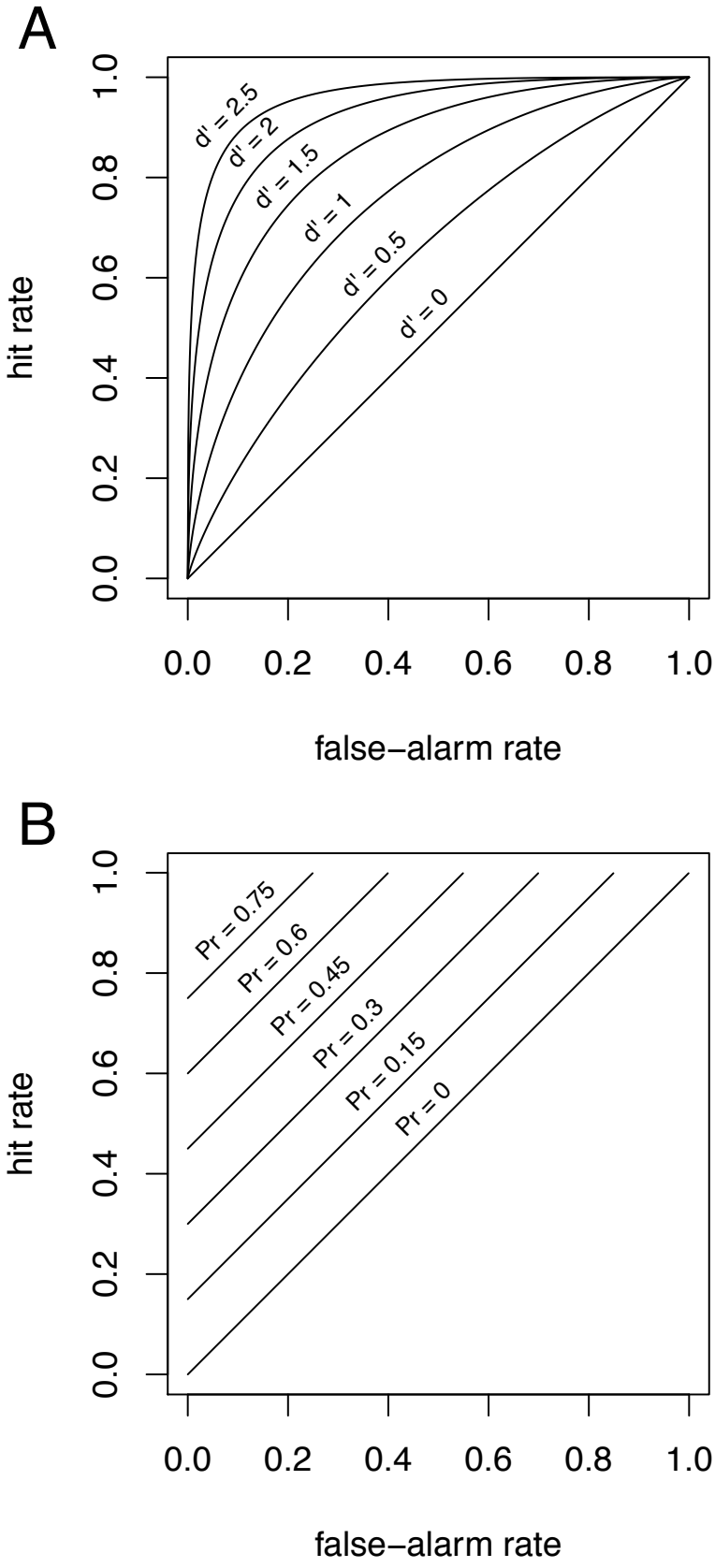

Figure 8.2: ROC curves predicted by (A) the Gaussian equal variance signal detection theory, and (B) two high-threshold theory ($P_o = P_n$).

8.2 Threshold Theory

In contrast to SDT which proposes a continuum of recognition 'strength', threshold theories propose that, at test, observers enter one of a handful of discrete states. Figure 8.3 shows the popular two-high threshold (THT) theory. According to this model when an old item is presented an observer has a specific probability, P_o, of detecting this, in which case they certainly respond *old*. On some proportion of trials $(1 - P_o)$, however, the observer does not enter the detect state and must guess. The probability that the observer correctly guesses that the item is old is given by, B_r. The probability of correctly responding old, then, is the sum of the branches ending in an *old* response, $h = P_o + (1 - P_o)B_r$. On trials where a new item is presented observers detect this at a rate of, P_n; therefore a false-alarm can only occur if the detect threshold is not passed and participants incorrectly guess. The probability of this occurring is given by, $f = (1 - P_n)B_r$. Assuming that the probability of entering the old detection state is the same as the probability of entering the new detection state ($P_o = P_n$, referred to as P_r: Snodgrass & Corwin, 1988) results in a simple measure of discriminability,

$$P_r = h - f, \tag{8.3}$$

which is commonly referred to as hit rate *corrected* for guessing (or corrected recognition). The probability that the observer guesses 'old' when they do not enter a detect state quantifies the bias in responding and is given by,

$$B_r = f/(1 - (h - f)), \tag{8.4}$$

(see, Snodgrass & Corwin, 1988). If the assumptions of the THT model are valid P_r provides an estimate of discriminibility detached from the observer's guessing bias.

This model is termed 'high threshold' as only old items can result in a state of old detection and only new items can elicit new detection. Low threshold models also exist in which observers can erroneously enter detect states (Luce, 1963b) however these are rarely considered (although see, Rouder, Province, Swagman, & Thiele, submitted). Also there is the one-high threshold model which is similar to the THT

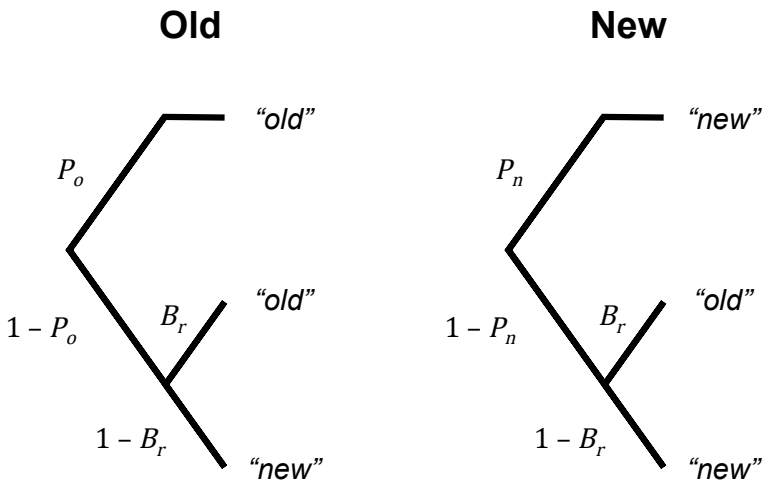

Figure 8.3: Multinomial processing tree model underlying the two-high threshold (THT) theory of detection/ recognition.

although states that observers cannot detect lures (i.e. $P_n = 0$). This results in the prediction that the probability of a false alarm depends solely on guessing, which does not accord with the observation that increases in hit rate are usually accompanied by decreases in false-alarm rate (see, Snodgrass & Corwin, 1988). Further, the simple processing models for estimating the number of items in VWM (k) from change detection tasks (e.g. Rouder et al., 2011) used throughout this thesis are extensions of THT theory. Indeed the tree like structure underlying this model is shown in Figure 8.3. Alternatively it is possible to think of threshold models in terms of detection theory in which the underlying distributions over the decision variable are rectangular (Macmillan & Creelman, 2005; Rotello et al., 2008).

It is interesting to note that the two-high threshold model of recognition can also be used to justify the use of proportion correct. As noted by Macmillan and Creelman (2005) the formula for proportion correct can be written as a linear function of hits minus false alarms, $p(c) = \frac{1}{2}(h - f) + \frac{1}{2}$. Thus using proportion correct to summarise performance (sensitivity) in a recognition task implicitly implies a THT decision model (Swets, 1986b), although this is rarely acknowledged.

Figure 8.2B shows the ROC predictions for the THT model at various threshold probabilities. The THT model considered here predicts linear ROCs with an intercept equal to P_r and slope of 1. Allowing the detect probabilities to differ (e.g.

$P_o > P_n$) results in a slope other than 1.

'Non-parametric' assessment of sensitivity

Green (1964) showed that threshold and detection theory were in agreement that the area under the ROC curve gives the expected proportion correct of an unbiased observer on a 2AFC task. Thus estimating the area under the ROC curve provides an assumption free estimate of sensitivity. Of course with a single hit and false-alarm rate pair per condition it is not possible to establish the shape of the ROC curve. However, Pollack and Norman (1964) proposed a method of approximating the area under the curve based on drawing triangles in the unit square ROC space. This approach is shown in Figure 8.4A. Drawing lines from (0, 0) and (1, 1) through the point (f, h) splits the square into 4 regions; one that would certainly be under a reasonable ROC curve (labelled I), one section that would fall above the curve (labelled S), and two sections in which the curve may fall (labelled $A1$ and $A2$). The measure proposed by Pollack and Norman (1964) is section I plus the average of sections $A1$ and $A2$ implied by the single (f, h) point: $A' = I + \frac{1}{2}(A1 + A2)$. It is not, however, the average of the largest and smallest areas implied by the single point (see, Smith, 1995; J. Zhang & Mueller, 2005), although this has not prevented its mass uptake. Hodos (1970) built on this and proposed a bias measure based on the relative areas of the two right-triangles formed by $A1 + S$ and $A2 + S$. Larger area in the $A1 + S$ triangle implies a tendency towards responding 'old', whereas greater area in the $A2 + S$ triangle implies a tendency towards answering 'new'. Grier (1971) and Aaronson and Watts (1987) provide the computing formulae for A' for above and below chance performance, respectively

$$\begin{aligned} A' &= \frac{1}{2} + \frac{(h - f)(1 + h - f)}{4h(1 - f)}, \quad h \geq f \\ A' &= \frac{1}{2} - \frac{(f - h)(1 + f - h)}{4f(1 - h)}, \quad h < f. \end{aligned} \tag{8.5}$$

With researchers concerned about the distributional assumptions of recognition measures, these ostensibly 'non-parametric' measures of sensitivity have gained widespread use in cognitive psychology. However, many have pointed out that just

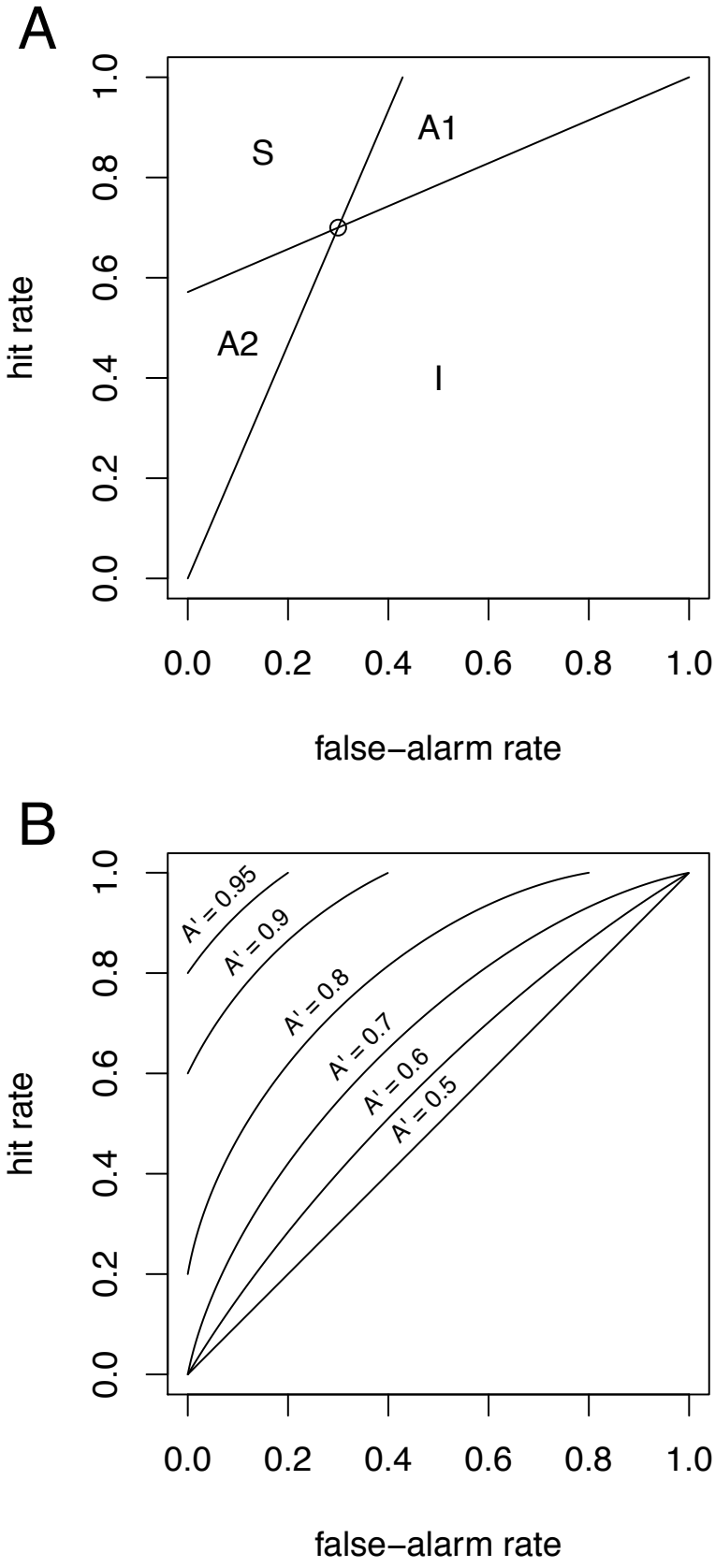

Figure 8.4: (A) Estimating area under the ROC curve with a single (f, h) pair. See text for description of the four segments. (B) the ROC curves implied by A'.

because these measures were derived without explicit reference to distributions it does not follow that they are non-parametric. In particular, Macmillan and Creelman have shown that the bias measure proposed by Hodos can be derived from a detection theory model assuming logistic distributions on the decision variable (Macmillan & Creelman, 1990). Further, in a later paper, they show that the underlying distributions implied by the measure A' change with the sensitivity of the observer (Macmillan & Creelman, 1996). When performance is high A' is consistent with a threshold model assuming (roughly) rectangular underlying distributions, whereas as performance lowers the assumptions increasingly reflect those of a logistic SDT model (see also Pastore et al., 2003; Macmillan & Creelman, 2005). This is also clear when the ROC curves implied by A' are plotted as shown in Figure 8.4B (see also, Pollack & Norman, 1964, Figure 2). Further, unlike the measures derived from detection and threshold theories, there is no underlying model that unifies measures of sensitivity and bias in the 'non-parametric' approach (Macmillan & Creelman, 1996; Pastore et al., 2003). Nevertheless A' is frequently used, especially in the literature on feature binding in VWM, therefore we included it in the present simulation study.

8.3 Previous Simulation Studies

Schooler and Shiffrin (2005) were the first to assess how the choice of measure could affect conclusions regarding sensitivity differences between experimental conditions. They were primarily interested in the efficacy of measures when the available data are sparse due to small trial numbers per participant. Their simulated data were generated using a GEV-SDT model in which two conditions either did or did not differ in terms of sensitivity, allowing them to assess type II and type I error rate, respectively. Provided the two conditions did not differ in terms of criterion placement the THT measure P_r performed quite well with type I error rates around the accepted value of 0.05. However, when conditions differ in criterion placement type I error rates for this measure were high and unsurprisingly d' performed better. When conditions truly differed in terms of sensitivity power was greatly improved by using

d' relative to P_r, again unsurprisingly as it matched the generative model.

Rotello et al. (2008) provided a more comprehensive set of simulation studies in which they generated data using an underlying SDT (Gaussian distributions) or THT (rectangular distributions) structure. They assessed the type I and type II (1 - power) error rates of repeated measures t-tests on multiple commonly used measures—including d', A' and proportion correct. Estimating sensitivity with a measure that did not match the generative model led to unacceptable type I error rates, provided that conditions differed in response bias. Type I errors were increasingly likely with bigger criterion differences between conditions and at higher levels of sensitivity. The error rate associated with A' was large regardless of the underlying distributions. They also conducted power analysis for their measures where there was true variation in sensitivity but fixed bias. All measures performed fairly well including A', especially with low overall sensitivity and small numbers of trials. However, given its unacceptably high type I error rates Rotello et al. (2008) council against the use of A' *in any situation*.

8.4 Rationale for the Present Simulations

The work of Rotello et al. (2008) and Schooler and Shiffrin (2005) clearly shows that, in the case of a comparison between two experimental conditions, a misguided choice of sensitivity measure can result in errors *provided the conditions differ in response bias*. The reason for this is clearly seen in the ROC plots depicted in Figure 8.2; if both the underlying bias and sensitivity of an observer are the same across two conditions then measures agree as both conditions imply the same point in ROC space. When bias is varied between conditions the same point in ROC space is no longer occupied by both conditions (i.e. they occupy different points on the same isosensitivity curve) and, therefore, the measures of sensitivity disagree.

One contribution of the present work is to point out that for *interactions* between experimental conditions it is possible for measures of sensitivity to disagree *without any variation in response bias*. The only condition that must be met is that each experimental factor should result in a main effect. The simplest situation to consider

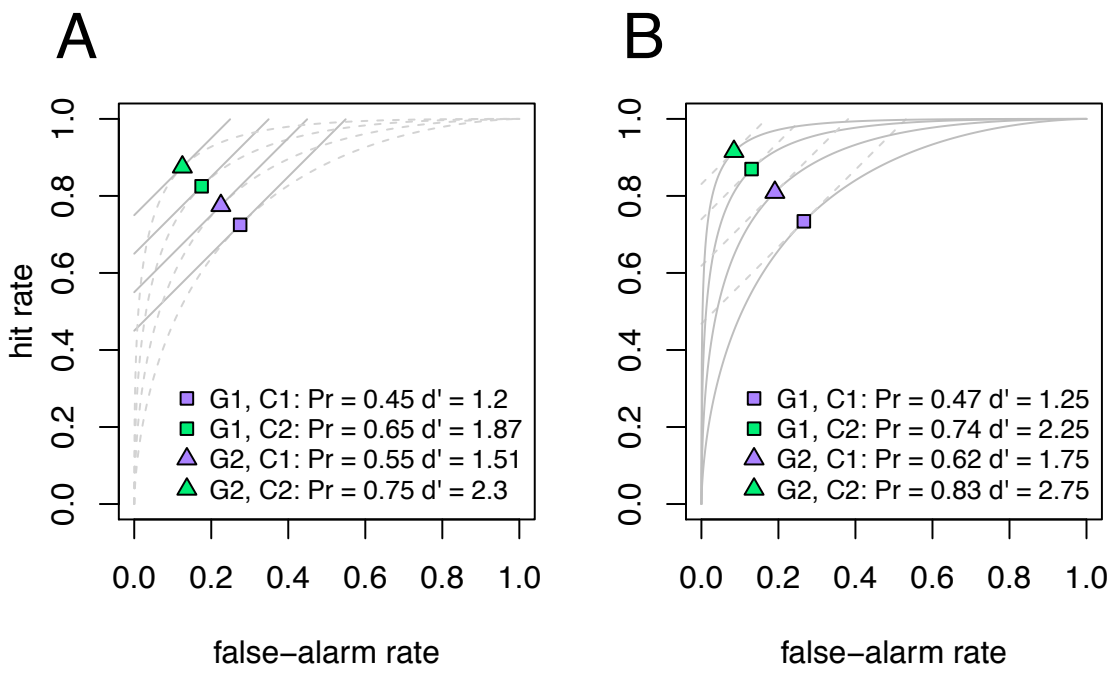

Figure 8.5: Plots showing that, when considering interactions, it is possible for THT and GEV-SDT measures to disagree without variation in bias. (A) non-interaction with P_r but interaction using d' (B) interaction using P_r but main effects only with d'. G1 and G2 refer to groups and C1 and C2 refer to conditions.

is a two-way design with two-levels to each factor; here we consider one between-subjects factor, *group*, and one within-subjects, *condition*. Figure 8.5 depicts two situations where, in absence of variation in response bias, P_r and d' would give opposing answers to the question of a group × condition interaction effect. Panel A presents the case where for P_r there are two main effects (conditions 1 and 2 differ by 0.2 and the groups differ by 0.1) but no interaction. When d' is calculated using the (f, h) pairs a clear interaction effect emerges where the effect of condition for group 1 is smaller (0.67) than it is for group 2 (0.79). The reason for this distortion can be seen in the ROC plots of Figure 8.2; at increasingly high levels of sensitivity (d') the curves become more compressed and thus the equally spaced points in ROC space shown in Figure 8.2A imply increasingly large values of d' resulting in an *overadditive* interaction.

Figure 8.2B presents the case where no interaction is present using d' (conditions 1 and 2 differ by 1 and the groups differ by 0.5) but an *underadditive* interaction is

present with P_r, such that the difference between conditions is larger in the lower performing group 1 (0.27) relative to group 2 (0.21). It is important to reiterate that this disagreement arises without any difference in response bias (all the points lie on the negative diagonal) and occurs when each experimental factor, in this 2×2 design, produces its own main effect.

Simulation allows us to assess the extent to which these issues cause problems in interpreting data for reasonable research designs. In addition to assessing the type I error rates of recognition measures in the absence of variation in response bias we also look at the effect of varying bias between groups and conditions. Previous simulation studies in a similar vein have assessed the effect of applying different corrections for hit and false-alarm rates of 0 or 1, however, we did not include this aspect in our simulations as this previous work suggested that the method applied had little effect (Rotello et al., 2008; Schooler & Shiffrin, 2005). The structure of our simulation studies is described next.

8.5 Structure of the Simulations

Simulated data sets contain two groups with N^S hypothetical subjects each performing in two conditions with N^T target present and N^T target absent trials per condition. Each trial is drawn from a Bernoulli distribution with the probability of success determined by the underlying model parameters for a given subject in a given condition. For both SDT and THT simulations the underlying parameters, p, are determined using the same general model:

$$p_{s,j,k} = \beta_0 + \beta_1 x_j + \beta_2 x_k + b_s,$$

where β_0 is the grand mean (for example, average true sensitivity, d') and β_1 and β_2 are deflections from the grand mean associated with the factors J (between subjects) and K (within subjects), respectively. x_j and x_k are indicator variables that are set to -1 if the observation comes from level 1 of the factor or to 1 for level 2 of the factor. Consequently positive values of our deflation parameters mean lower parameter values at level 1. The final component, b_s, reflects random variation in

the parameter value due to subject, s. Note that this only affects the overall level of performance between hypothetical participants and there is no variability associated with the effect of condition. This variation was specified slightly differently for the SDT and THT simulations as will be outlined later.

Both sensitivity and bias parameters were determined using this linear equation. Specifying the parameters in this way allows us to manipulate the overall level of sensitivity (or bias), via β_0, and the magnitude of the group and condition effects, via β_1 and β_2, (as well as the number of subjects per group and trials per condition) with no interaction present in the underlying model. The parameters for each hypothetical participant are used to generate expected hit and false-alarm rates according to the assumptions of the SDT and THT models; the specifics are outlined below.

Gaussian equal variance simulations

Here the crucial underlying parameters were sensitivity (d') and criterion placement (c). Both were determined by the linear model described above although to distinguish the parameters from those described below we refer to the sensitivity parameters $\{\beta_0^{d'}, \beta_1^{d'}, \beta_2^{d'}, b_s^{d'}\}$ and the criterion parameters $\{\beta_0^{c}, \beta_1^{c}, \beta_2^{c}, b_s^{c}\}$. The β parameters were deterministic whereas b_s was drawn from a normal distribution with a mean of 0 and standard deviation $\sigma^{d'}$ or σ^{c} for sensitivity and criterion, respectively. These parameters determined the expected hit and false-alarm probabilities for each subject, group, condition combination as described in the explanation of the GEV-SDT model above.

Two high threshold simulations

According to THT theory the crucial parameters are the probability that the relevant information crosses the thresholds (P_r) and the bias towards responding *old* (or *target present*) when the relevant information is not present (B_r). These parameters were set using the linear equation above with discriminability parameters $\{\beta_0^{P_r}, \beta_1^{P_r}, \beta_2^{P_r}, b_s^{P_r}\}$ and bias parameters $\{\beta_0^{B_r}, \beta_1^{B_r}, \beta_2^{B_r}, b_s^{B_r}\}$. The random subject ef-

fect was achieved by taking $2N^S$ draws from the normal cumulative distribution with a mean of 0 and standard deviation σ_{P_r} or σ_{B_r} for discriminability and bias, respectively and subtracting the resulting value from 0.5 ($b_s^{P_r} \sim 0.5 - \Phi(\text{Normal}(0, \sigma_{P_r}))$). While this may seem convoluted, specifying subject variability in this way means that variability can be set on a similar scale to the above Gaussian simulations. Unlike d' and c, P_r and B_r are constrained to lie between 0 and 1, so values greater than 1 were rounded down to 1 and less than 0 rounded up to 0. Expected hit and false-alarm rates for each subject, group, condition combination were computed as per the description of the THT model given above.

For both the Gaussian and threshold simulations the expected hit rate from the underlying parameters was used to set the underlying probability of success for N^T Bernoulli trials representing 'old' (or target present) trials. Similarly the expected false-alarm rate was used to set the probability of success for N^T 'new' (or target absent) Bernoulli trials. The simulated hit and false-alarm rates were then used to calculate d', P_r, and A' using Equations 8.1, 8.3, and 8.5, respectively (see also, Macmillan & Creelman, 2005; Snodgrass & Corwin, 1988; Stanislaw & Todorov, 1999). As d' is undefined when either the hit or false-alarm rate is equal to 1 or 0 we used the commonly used correction of adjusting rates of 1 down to 0.99 and rates of 0 to 0.01. The other measures are able to handle perfect performance—and this is occasionally cited as justification for their use—so the adjustment was not made for these measures.

These estimates were then analysed with a two-way mixed ANOVA, using the `R` function `aov`, to obtain a p value for the crucial Group $\times$ Condition interaction. For each simulation this procedure was repeated 1000 times to obtain the expected interaction type I error rate for the given measure and parameter settings.

8.6 Simulation Study 1: Error Rate Without Variation in Bias

For the first set of simulations we assessed the extent to which type I errors occur for Group by Condition interaction effects with simulated data in which there was no interaction effect present for sensitivity and no variation at all in response bias. This study also assessed how several factors modulate errors rates, to do so we manipulated; 1) overall sensitivity (via the grand mean parameter, β_0), 2) the magnitude of the main effects of group and condition (β_1 and β_2), and 3) the sample size in terms of the number of participants per group (N^S) or number of trials per condition (N^T; in separate simulations).

Thus for the GEV and THT simulations we selected 5 grand mean values and 3 main effect sizes (as well as no effect)—in this case fixing the two main effects to be of the same magnitude to reduce the number of simulations needed—and simulated data for 12, 24, or 48 hypothetical participants per group or trials per condition. In this case we were not interested in the effects of varying response bias, therefore bias parameters were set to zero for all simulations (except for $\beta_0^{B_r}$ which was set to 0.5).

Parameters were selected to cover a wide range of possible values according to the underlying model of recognition. For the GEV simulations grand mean parameters ($\beta_0^{d'}$) ranged from 1.5 to 3.5 in steps of 0.5 and deflection parameters were 0.125, 0.25, and 0.5 ($\beta_1^{d'} = \beta_2^{d'}$) for small, medium, and large effects, respectively. Of course these values, and their verbal labels, are arbitrary but allow us to cover a range of sensitivities. These parameters imply that the lowest expected sensitivity will be 0.5 for Group 1 in Condition 1, whereas when the grand mean is 3.5 the expected sensitivity of Group 2 in Condition 2 is 4.5. Although d' can theoretically range from 0 to ∞ a reasonable maximum for d' is approximately 4.65 which (if $c = 0$) produces expected hit and false-alarm rates of 0.99 and 0.01, respectively.

For the THT simulation mean levels of P_r ranged from 0.4 to 0.8 in steps of 0.1 and deflection effects were 0.025, 0.05, and 0.1 for the small, medium, and large

effects, respectively. Thus the lowest expected value was 0.2 for Group 1 in Condition 1 and the highest was 1 for the second group in Condition 2 (i.e. a ceiling effect).

Results

Gaussian Equal Variance SDT

Figure 8.6 presents the estimated type I error rates (number of significant interactions divided by 1000) for d', A', and P_r with a SDT underlying model with equal variance Gaussian evidence distributions. A couple of noteworthy patterns stand-out; firstly error rates for all measures are at or around the conventionally accepted rate of 0.05 when there are no main effects or the true main effect of group and condition (fixed to be the same magnitude) is small. When there are fairly large main effects the error rates for A' and P_r depart from accepted levels and increasing the number of participants per group exacerbates this (Figure 8.6 top to bottom panels). Of the measures, A' clearly is more likely to erroneously give evidence for an interaction effect where none exists.

The effect of increasing the overall mean level of sensitivity is slightly more complex. Using P_r on GEV-SDT data becomes increasingly problematic as the underlying sensitivity increases, whereas for A' errors become somewhat less pronounced as sensitivity increases. In both cases however when there are large main effects the type I error rate is unacceptable (0.705 for P_r and 0.856 for A'). The principled measure in this case, d', also exhibits exacerbated error rates when underlying sensitivity is high and there are large main effects. This is clearly due to a ceiling effect which restricts performance of Group 2 in the second Condition (where expected $d' = 4.5$).

Thus it is clear that interaction type I error rates for measures inconsistent with the generative model can arise with no variation in response bias. The necessary condition is that there are clear main effects of the experimental factors which, as explained above produces 4 points in ROC space and consequently disagreement between different measures. This is not a trivial problem with error rates occasionally well in excess of 0.5.

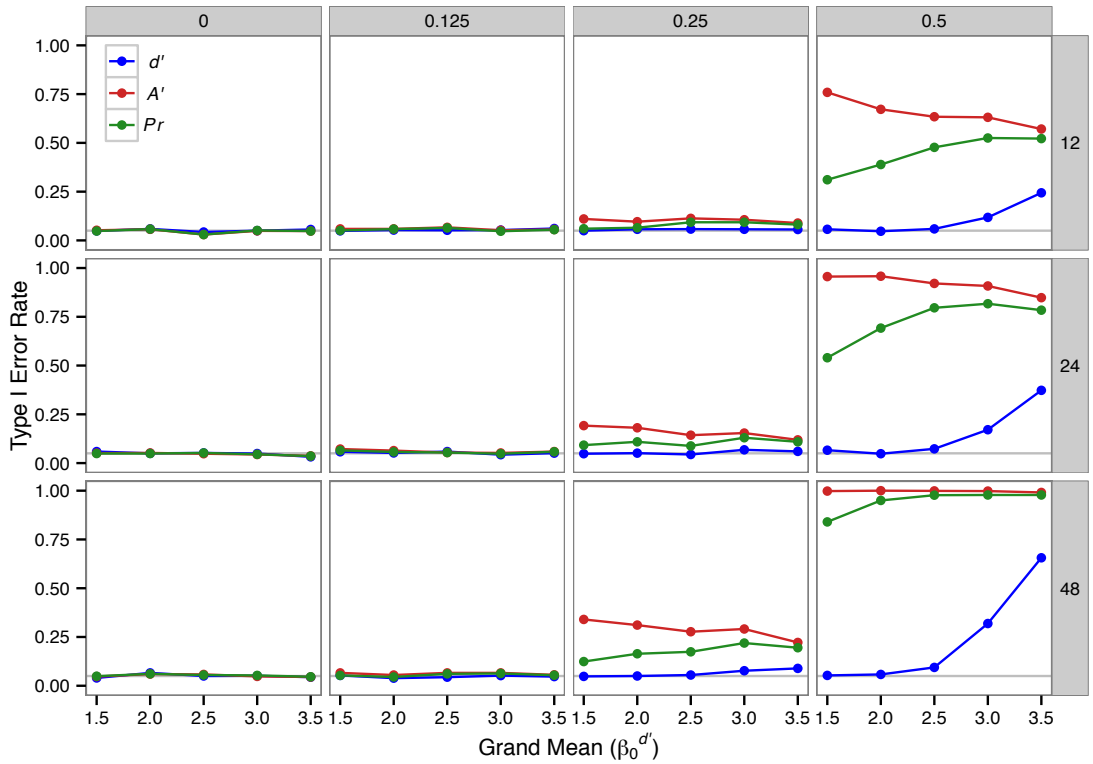

Figure 8.6: Simulation 1: Type I error rates with a GEV-SDT underlying model. Panels from left to right show the effects of increasing the size of the true underlying main effects ($\beta_1^{d'} = \beta_2^{d'}$). Panels from top to bottom show the effect of increasing sample size per group.

In line with the simulations of Schooler and Shiffrin (2005) and Rotello et al. (2008), when there was no true effect of either Group or Condition the type I error rate for tests of a main effect was 0.05 regardless of the measure used. When there was a true effect power for the effect of Group was generally lower than the effect of Condition, due to the added subject variability in underlying sensitivity. In the 'small' effect condition ($\beta_{1,2}^{d'} = 0.125$) the effect of Condition is correctly detected around 65% of the time regardless of measure used, whereas the effect of Group is detected around 45% of the time. In this case the type I error rate for the interaction is no greater than 0.06. It is only when power for the main effects exceeds 0.8, in the 'medium' effect condition (power = 0.87 for Group and 0.95 for Condition), that type I error rates for A' and P_r depart from accepted levels (0.18 and 0.12, respectively). When the true main effects were 'large', and power is as good as 1 (≥ 0.998) for both main effects, the interaction type I error rate is uncontrollable

(0.856 for A' and 0.705 for P_r).

Two-High Threshold

As shown in Figure 8.7, type I error rates for the simulation with a THT generative model remain under control until there are large main effects of Group and Condition. As shown in the rightmost panels, error rate for both A' and d' depends on the overall level of discriminability ($\beta_0^{P_r}$). The frequency of errors is more pronounced with d' relative to A' until the very highest grand mean sensitivity is reached, at which point the error rate for d' drops to a similar level as P_r. As noted above, erroneously applying d' to data conforming to the predictions of a THT model leads to the impression that sensitivity differences between conditions are *larger* in groups that are more sensitive overall. The ceiling effect encountered at high underlying levels of P_r quashes this *overadditive* tendency. The effect of ceiling discriminability is also seen when the correct measure, P_r, is applied (see the rightmost panels of Figure 8.7).

Again, in line with previous simulation work (Rotello et al., 2008; Schooler & Shiffrin, 2005), when there was no true effect of either Group or Condition the significance rate for these components was at the expected rate of 0.05. When there were true orthogonal main effects, power to detect these effects was similar across measures. For our 'small' effect, power was around 25% for the between subjects effect and around 59% for the within subjects effect and for the 'medium' size effect these values were 65% and 95%, respectively. In both of these cases the error rate associated with choosing an inappropriate measure did not diverge from 0.05. However, when both effects were 'large' power for the main effects was greater than 96% and erroneous interactions were frequent with both A' (0.242) and d' (0.295). Thus, as with the GEV-SDT simulation, unacceptable type I error rates for interactions appear to be dependent on sufficient power to detect *both* main effects. We directly assess this in our next series of simulations.

For both the GEV-SDT and THT simulations increasing the number of participants per group had the effect of increasing the overall error rate but did not appear

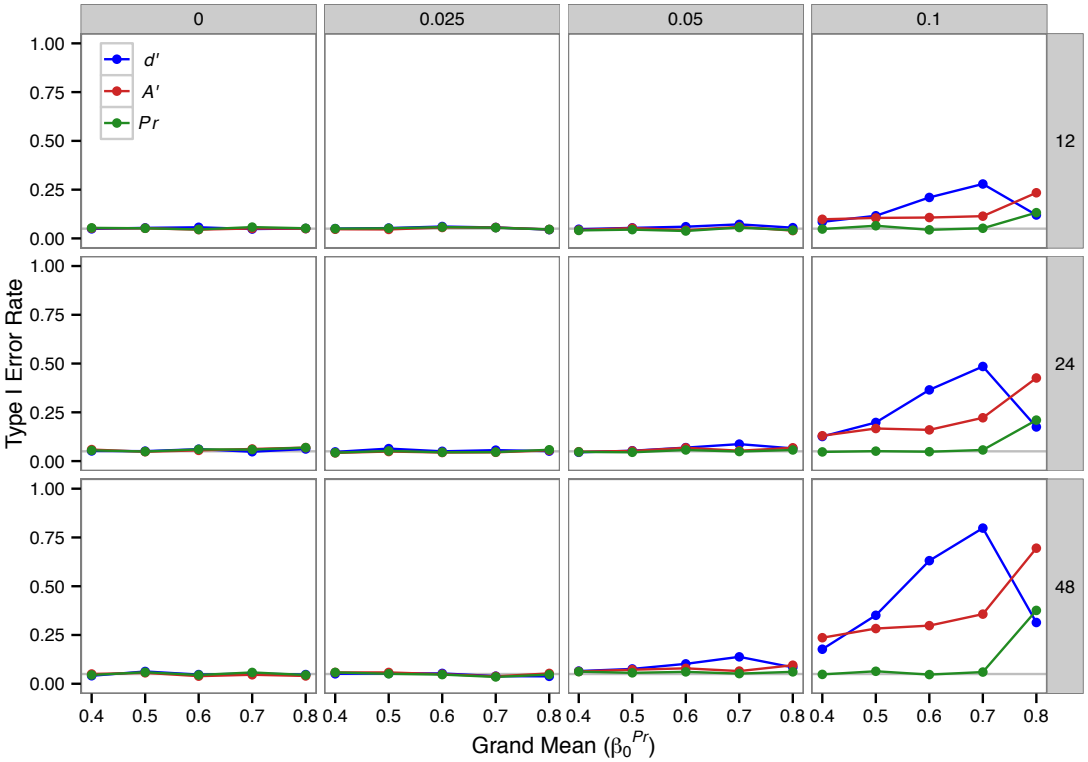

Figure 8.7: Simulation 1: Type I error rates with a THT underlying model. Panels from left to right show the effects of increasing the size of the true underlying main effects ($\beta_1^{Pr} = \beta_2^{Pr}$). Panels from top to bottom show the effect of increasing sample size per group.

to modulate the effect of varying other parameters. Increasing the number of trials in each condition was found to have an almost identical effect, so the results of that simulation are not presented here.

8.7 Simulation Study 2: Orthogonally Varying Main Effects

Simulation 1 showed that applying an unprincipled measure of recognition performance leads to inflated type I error rates for a Group by Condition interaction effect. This occurred in the absence of variation in response bias provided that each factor produced a medium to large main effect on sensitivity. However, in those simulations the main effects were constrained to be the same size. In the present series of simulations we orthogonally vary the size of effects, selecting from the same

values used in the previous study (GEV-SDT simulations $\{0, 0.125, 0.25, 0.5\}$, THT simulations $\{0, 0.025, 0.05, 0.1\}$). For all subsequent simulations sample size, both in terms of number of trials per condition (N^T) and number of subjects per group (N^S) was fixed to 24, which are fairly representative of studies in the cognitive ageing literature.

Results

Gaussian Equal Variance SDT

Figure 8.8 presents type I errors when main effect sizes are orthogonally manipulated (left to right = Group effect, top to bottom = Condition effect) using a GEV-SDT generative model. The general patterns observed in the first simulations are observed here; error rates for A' are far more pronounced than P_r and both depend on the grand mean underlying sensitivity ($\beta_0^{d'}$). Error rates for d' become problematic when expected sensitivity is especially high. It is clear from Figure 8.8 that error rates depart from 0.05 when at least one main effect is fairly large and the other effect is non-zero. There are no clear differences contrasting above and below the diagonal, suggesting that increasing the Group effect has a similar effect to increasing the (more reliable) Condition effect.

Two-High Threshold

For the simulations with a THT underlying structure, when the Group effect was 0.05 and the Condition effect was 0.1 the error rate for d' rose to approximately 0.17 (see Figure 8.9). This was similar when the Group effect was 'large' and the condition effect was 'medium' where the d' error rate rose to approximately 0.2. It is important to note that in this case the power to detect these main effects was essentially 1 (0.998 and 0.999 respectively). Of course the most pronounced error rates were observed when both main effects were large.

These simulations show that *both* main effects are needed to cause inflated type I error rates for the two-way interaction test. This is most pronounced when one effect is quite large and the other is at least medium in size.

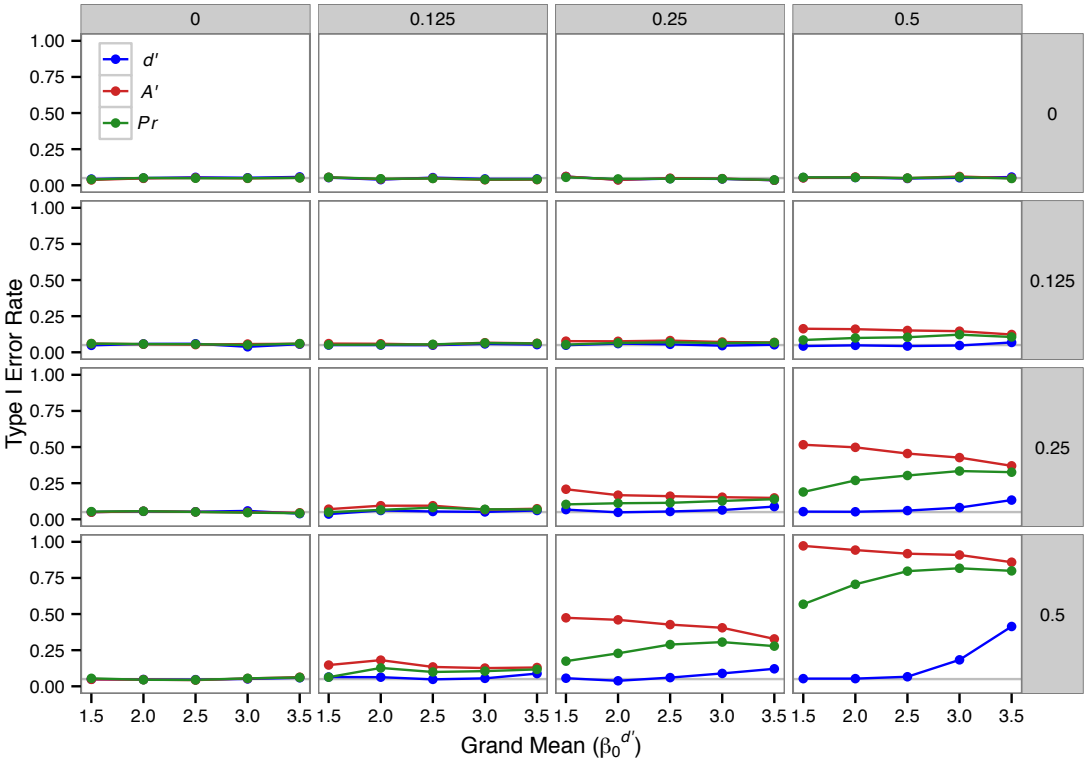

Figure 8.8: Simulation 2: Orthogonally varying main effect sizes with a GEV-SDT underlying model. Panels from left to right show the effects of increasing the magnitude of the underlying Group effect ($\beta_1^{d'}$). Panels from top to bottom show the effect of increasing the magnitude of the underlying Condition effect ($\beta_2^{d'}$).

8.8 Simulation Study 3: Varying Overall Response Bias

So far we have been primarily interested in assessing type I error rates in the absence of variation in response bias. The first two studies clearly show that, provided there are sufficient main effects in the design, false-positives for interaction effects can become commonplace if an unprincipled measure is applied. In the third set of simulations we assessed the effect of introducing variation in response bias. For this simulation the variation was achieved by allowing our hypothetical subjects to vary in response bias as well as modifying the grand mean bias exhibited. In this case we were not interested in varying bias by either group or condition (see Simulations 4 and 5 below).

For the GEV-SDT simulations grand mean criterion placement, β_0^c, was selected

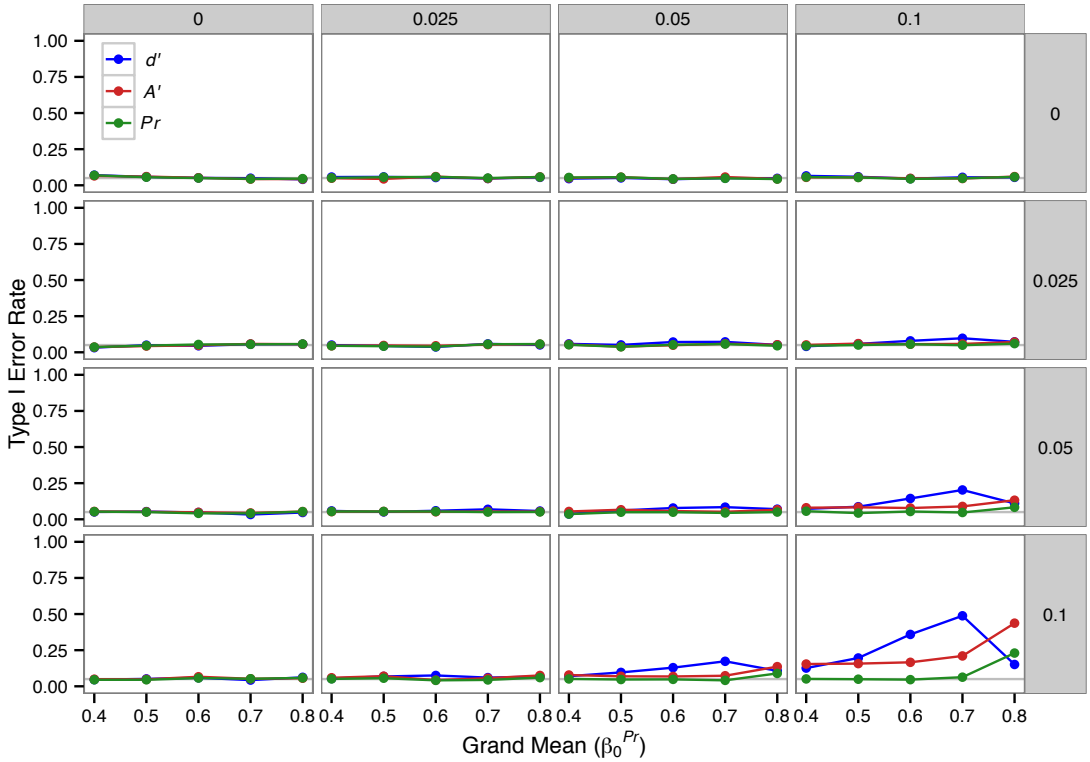

Figure 8.9: Simulation 2: Orthogonally varying main effect sizes with a THT underlying model. Panels from left to right show the effects of increasing the magnitude of the underlying Group effect (β_1^{Pr}). Panels from top to bottom show the effect of increasing the magnitude of the underlying Condition effect (β_2^{Pr}).

from 4 values from liberal responding to conservative, $\{-1, -0.5, 0, 0.5, 1\}$. Each subject's criterion placement was also determined by the random effect with a mean of 0 and standard deviation, $\sigma^c = 0.3$. For the THT simulations grand mean biases, β_0^{Br}, were selected from $\{0.3, 0.4, 0.5, 0.6, 0.7\}$, thus from a tendency to say 'new' when failing to enter a detect state towards a tendency to say 'old'. The random subject parameter, σ^{Br}, was also set to 0.3.

Results

Gaussian Equal Variance SDT

Figure 8.10 shows the effect of varying overall response bias with a GEV-SDT underlying model. It is clear that varying overall bias is not sufficient to cause serious interaction type I errors when the true main effects on sensitivity are rather small.

When sensitivity effects are large enough to cause errors the effect of varying criterion in either direction is to *reduce* the error associated with choosing an unprincipled measure (see panels left to right). At the extreme end of our criterion values the type I error for P_r is at an acceptable level when the true sensitivity effect is moderate. When main effects on sensitivity are large criterion placement also has a pronounced effect on the relationship between error rate and overall sensitivity ($\beta_0^{d'}$). For P_r the relationship is weakened whereas for A' the error rate becomes increasingly dependent on overall sensitivity. The reason for this change can be understood with reference to the GEV-SDT isosensitivity curves presented in Figure 8.2A. As criterion departs from the negative diagonal the curves implied by different sensitivities are forced closer together, making the (f, h) pairs predicted by d' and P_r under no-interaction more similar.

The bottom panels of Figure 8.10 show that, when sensitivity effects are large, type I error rate for d' can be inflated. Again this is caused by ceiling or floor effects in hit and false-alarm rate. At the extreme ends of criterion placement (-1, 1) error rates are highest at low overall sensitivity where as at more moderate criterion placement (-0.5, 0, 0.5) errors are increasingly common at higher sensitivity.

Two-High Threshold

In contrast to the GEV-SDT simulations above the simulations with a THT underlying model reveal little effect of varying the bias towards responding 'old' when a detect state is not reached (see Figure 8.11). The error rate associated with d' does not rise to quite the same level at more extreme levels of bias but otherwise the bottom panels of Figure 8.11 are quite similar. Once again a key to interpreting this can be found in the THT ROC, shown in Figure 8.2B. Varying bias moves predictions for (f, h) pairs at a given discriminability along the diagonal line but, as the lines are parallel, does not modulate the difference between different threshold probabilities. Thus the predictions of THT and GEV-SDT still greatly disagree.

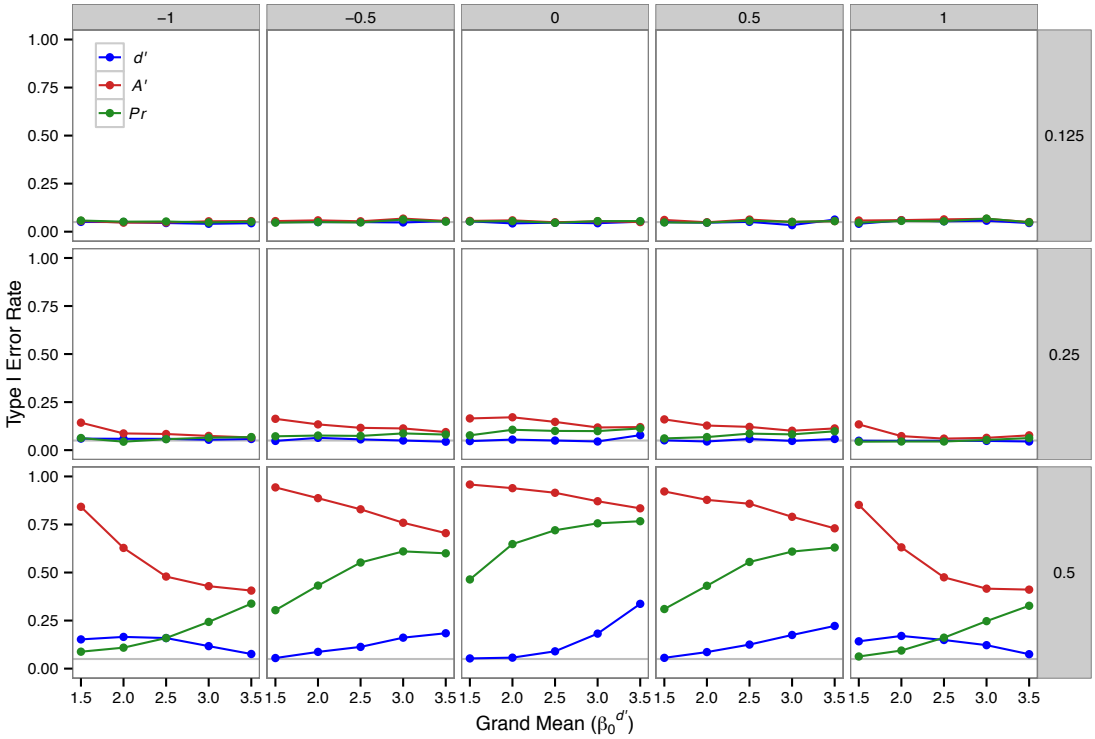

Figure 8.10: Simulation 3: Error rates for GEV-SDT data when varying grand mean criterion placement as well as introducing random subject variability in criterion setting. Panels from left to right show the effect of manipulating the overall bias exhibited by observers in criterion placement (β_0^c). Panels from top to bottom show the effects of increasing the magnitude of the underlying group effect ($\beta_1^{d'} = \beta_2^{d'}$).

8.9 Simulation Study 4: Varying Bias Between Groups and Conditions

The previous simulation allowed response bias to differ from a neutral position and allowed individuals to differ in their tendency to favour one response option. However, there was no true effect of either Group or Condition on bias. It is conceivable that different groups differ in their response tendencies and that an experimental manipulation, as well as affecting sensitivity, may also have an effect on bias. Therefore, in the present simulation we assessed the effect of introducing true main effects on response bias. To do this, without having an unwieldy number of simulations, we fixed the sensitivity grand mean to 0.6 for the THT simulations (β_0^{Pr}) and to 2 for GEV-SDT simulations ($\beta_0^{d'}$). The magnitude of the two true main effects on sensitivity was also fixed ($\beta_1^{Pr} = \beta_2^{Pr} = 0.1$ and $\beta_1^{d'} = \beta_2^{d'} = 0.25$). Note that here there

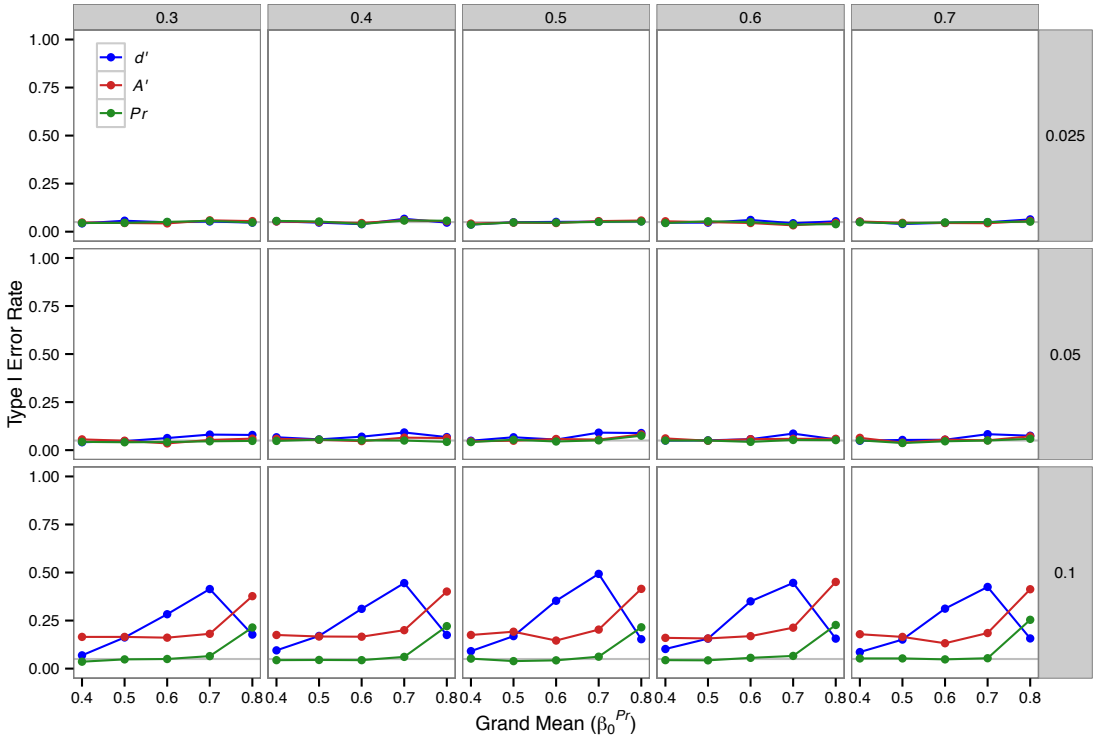

Figure 8.11: Simulation 3: Error rates for THT data when varying grand mean criterion placement as well as introducing random subject variability in criterion setting. Panels from left to right show the effect of manipulating the overall bias exhibited by observers in criterion placement (β_0^{Br}). Panels from top to bottom show the effects of increasing the magnitude of the underlying group effect ($\beta_1^{Pr} = \beta_2^{Pr}$).

are true main effects of Group and Condition on sensitivity. In the final simulation we take the logical step of assessing interaction errors with variation in bias but *no* variation in sensitivity.

The main effects on response bias were selected from a range of values, including no effect. For the GEV-SDT simulations β_1^c and β_2^c were selected from $\{0, 0.125, 0.25, 0.5\}$ and for the THT simulations β_1^{Br} and β_2^{Br} were selected from $\{0, 0.025, 0.05, 0.1\}$.

Results

Gaussian Equal Variance SDT

Figure 8.12 presents the simulation results when both group and condition were allowed to very in criterion placement using a GEV-SDT architecture. Focusing on

the panels along the top and left of the Figure shows the situation where there is a *single* main effect on response bias. In this case the only appreciable effect is on A' in that at more extreme main effect sizes error rate become increasingly dependent on the overall level of bias exhibited by observers. Error rates rarely exceed 0.25.

The remaining 9 panels depict the case where both Group and Condition have an effect on response bias. An interesting pattern emerges in that the error rate for A' and especially for P_r becomes increasingly like an inverted 'U' shape with more extreme errors when the average criterion (β_0^c) is neutral. Type I error rates drop to 0.05 as criterion becomes more liberal and do not show an equivalent drop for more conservative responding. d' also shows unacceptable error rates with large bias effects and a general trend towards conservative responding.

The reasons behind the asymmetrical shape of error rates as a function of mean response bias are complicated and of secondary importance. Briefly, the inverted 'U' shape arises due to the fact that the curvature of the ROC is largest when overall responding is neutral, making the difference between groups and conditions clearer and hence the disagreement between measures more pronounced. The asymmetry arises due to our parameter settings; using positive values for both sensitivity main effects and bias main effects results in more sensitive groups/ conditions exhibiting more conservative responding. This results in a pronounced floor effect—as the most sensitive conditions will produce few if any false-alarms—causing errors for all measures, including d'.

Two-High Threshold

The pattern observed with a THT generative model, which is presented in Figure 8.13, is much simpler. Allowing Groups and Conditions to differ in guessing bias increases the type I error for interactions when using d' but decreases the error associated with using A' (see Figure 8.13). As the size of the underlying main effects on bias increases, error rates for both measures are increasingly dependent on the average probability of guessing old ($\beta_0^{B_r}$). For A', error rates increase as the tendency to guess old increases whereas for d' errors are more likely when observers

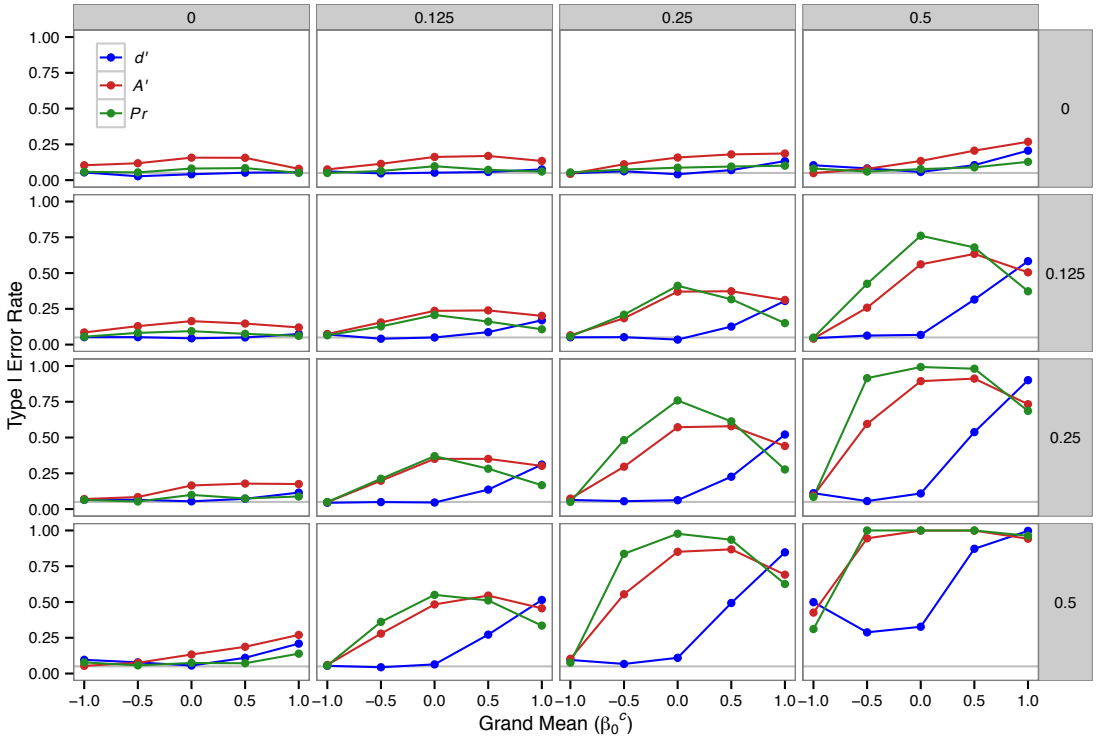

Figure 8.12: Simulation 4: The effect of varying criterion placement by group and condition with a GEV-SDT underlying model. Panels from left to right show the effect of increasing the effect of group on response bias (β_1^c). Panels from top to bottom show the effect of increasing the effect of condition on response bias (β_2^c). Note that the x-axis now describes the grand mean criterion placement (β_0^c).

tend to guess new. The error rate for P_r never departs from 0.05.

8.10 Simulation Study 5: Varying Bias and Not Sensitivity

The final simulation was essentially identical to Study 4 with the exception that deflection parameters for main effects on sensitivity were set to zero ($\beta_1^{Pr} = \beta_2^{Pr} = \beta_1^{d'} = \beta_2^{d'} = 0$). Therefore, this simulation assessed whether, in the absence of any sensitivity differences, type I errors arise for interactions when there are Group or Condition differences in bias.

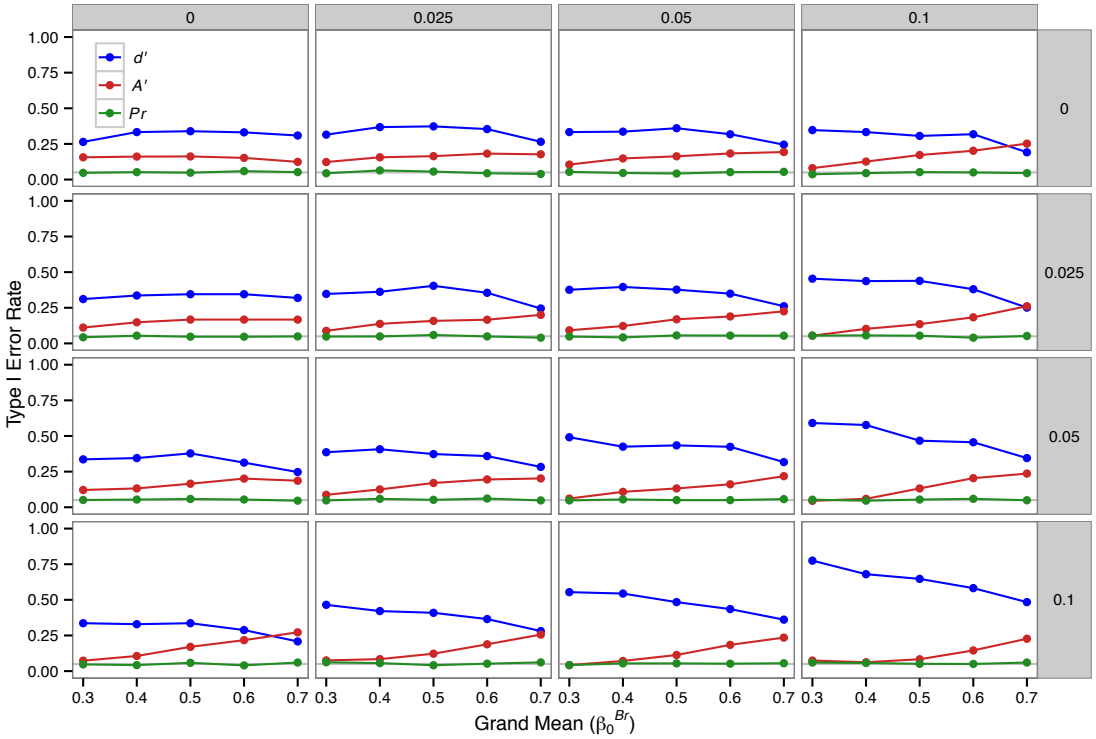

Figure 8.13: Simulation 4: The effect of varying response bias by group and condition with a THT underlying model. Panels from left to right show the effect of increasing the effect of group on response bias ($\beta_1^{B_r}$). Panels from top to bottom show the effect of increasing the effect of condition on response bias ($\beta_2^{B_r}$). Note that the x-axis now describes the grand mean bias ($\beta_0^{B_r}$).

Results

Gaussian Equal Variance SDT

Figure 8.14 presents the pattern of error rates arising from our GEV-SDT simulations. As in Study 4, a clear inverted 'U' shape can be observed in the interaction error rates for P_r and A' when main effects on criterion are medium-to-large. The cause of this shape is the same as that in Simulation 4, whereas in this case the 'U' shape is symmetrical as there were no main effects on sensitivity.

Two-High Threshold

Figure 8.15 shows the effect of allowing main effects of Group and Condition on guessing bias when there are no true effects on discriminability. Compared to the GEV-SDT simulations the pattern of results is very simple. Error rates do not

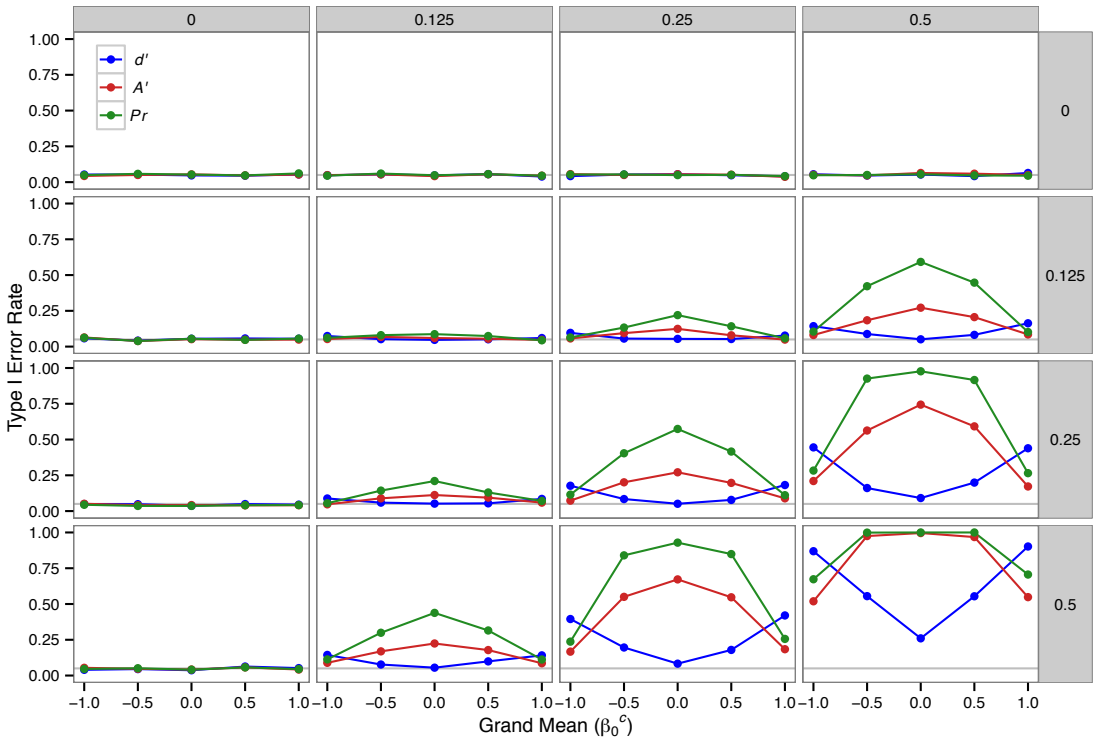

Figure 8.14: Simulation 5: The effect of varying criterion placement by group and condition with a GEV-SDT underlying model and no effects on sensitivity. Panels from left to right show the effect of increasing the effect of group on response bias (β_1^c). Panels from top to bottom show the effect of increasing the effect of condition on response bias (β_2^c). Note that the x-axis now describes the grand mean criterion placement (β_0^c).

greatly deviate from 0.05 until both bias effects are large. When groups and conditions both differ in guessing bias the error associated with choosing d' raises to approximately 0.15 and does not depend greatly on the overall bias exhibited by observers. Thus, in the absence of variation in detection probabilities, there appears to be little chance of a type I error when an inappropriate measure is applied to data conforming to THT expectations.

8.11 Discussion

Recognition performance summarised in terms of hit and false-alarm rate confounds the observer's ability to perform the discrimination with their inherent bias towards one of the response options. In order to separate the contribution of these two factors

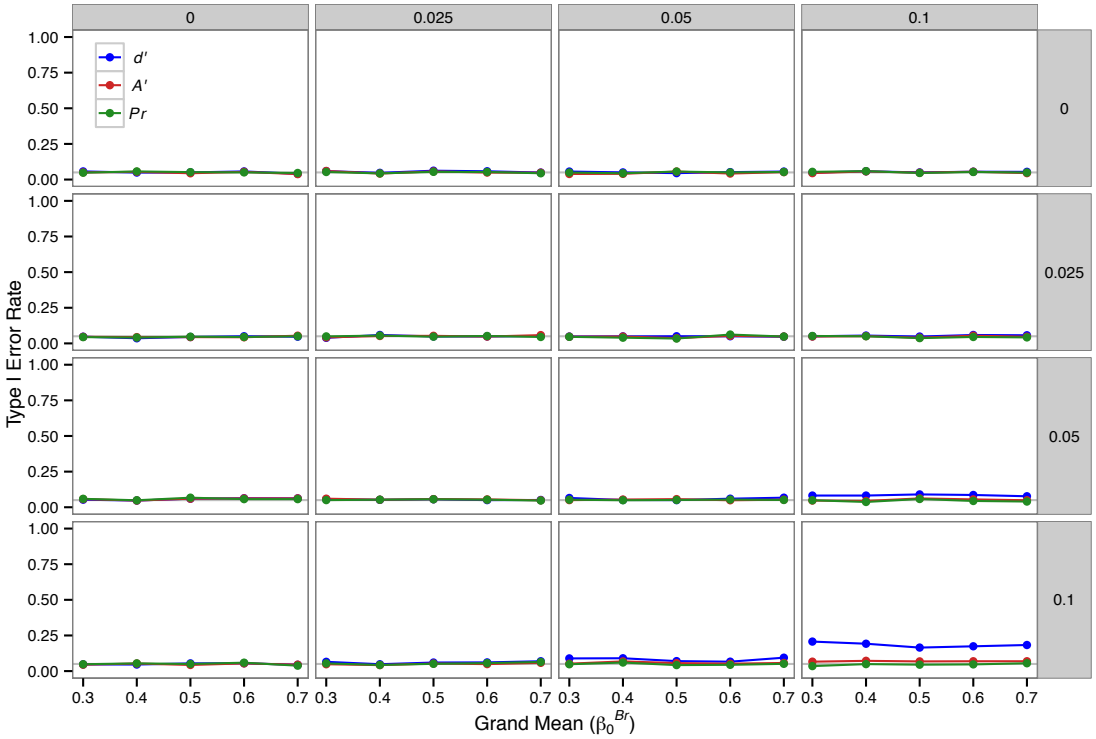

Figure 8.15: Simulation 5: The effect of varying response bias by group and condition with a THT underlying model and no effects on discriminability. Panels from left to right show the effect of increasing the effect of group on response bias ($\beta_1^{B_r}$). Panels from top to bottom show the effect of increasing the effect of condition on response bias ($\beta_2^{B_r}$). Note that the x-axis now describes the grand mean bias ($\beta_0^{B_r}$).

a model of the underlying decision process is needed. The commonly used measures d' and P_r come from two broad classes—those that assume a continuous decision variable with observers comparing sampled values to a criterion and those assuming a handful of discrete states—and are particularly restricted realisations of these models. Previous work has established that, if an unprincipled measure is applied, two conditions may appear to differ in performance (i.e. sensitivity) where no true difference exists *provided that* they differ in response bias (Rotello et al., 2008; Schooler & Shiffrin, 2005). Here we have shown that erroneous conclusions can arise regarding interactions between variables in the absence of any difference in response bias (Simulation 1). Provided both group and condition, in this case, result in moderate to large size effects themselves the probability of erroneously encountering a significant interaction effect is quite large when an unprincipled measure is applied (Simulations 1 and 2). Variation in response bias between groups and conditions can

further compound this problem (Simulations 3 and 4) and can cause inflated type I error rates in the absence of any sensitivity differences (Simulation 5), particularly when the data conform to the expectations of a GEV-SDT model. Another clear finding from the present simulations is that, even when the chosen measure matches the underlying generative process, type I errors can become unmanageable when sensitivity is too high overall in a particular group. This was especially true for d' given the correction that must be made for hit or false-alarm rates of 0 or 1.

This should be of concern to researchers across the fields of cognitive ageing and neuropsychology more generally. Groups (e.g. young and old) typically differ in performance and it is often difficult to match the demands of theoretically interesting conditions. Thus researchers and practitioners alike should be skeptical of a finding suggesting that recognition (or detection) performance is disproportionately affected in a certain group under certain conditions unless the measure applied is sufficiently justified. For our present purposes we were interested in the effect the choice of measure has on type I error rate; however, another implication of Figure 8.5 is that it is possible to have a *true* interaction that appears merely as additive effects (i.e. is missed) when an unprincipled measure is applied. Thus it will be important for future work to assess the *power* of measures to detect interactions where they do exist. This is beyond the scope of the current work and, given that greater emphasis is typically placed on the avoidance of type I error rates, the following recommendations can be made.

Recommendations

Avoid ceiling performance

Avoiding perfect, or near perfect, performance is, of course, easier said than done and is usually a concern for researchers anyway. The present simulations reinforce this concern by clearly showing that high levels of performance can cause issues even when a principled measure is chosen. This is particularly the case when one group performs particularly well in one condition. In the present work we considered situations where sensitivity was above chance but the same problems will arise with

low performance.

Don't use A'

The clearest recommendation that can be made is *not to use* A'. Indeed this suggestion could be made without the present simulations and has been made multiple times (e.g. Rotello et al., 2008); however, given the pervasive use of this metric it deserves to be restated. A', and associated bias measures, *do* imply certain underlying evidence distributions and therefore should not be considered 'non-parametric' (Macmillan & Creelman, 1990, 1996; Pastore et al., 2003). Further, A' does not reflect the average of the minimum and maximum ROC areas implied by a single (f, h) point (Smith, 1995; J. Zhang & Mueller, 2005) and cannot be justified on this basis. The present work adds to the list of reasons to avoid A' by demonstrating that, across all the simulations conducted, it is the most likely to produce spurious significant interaction effects (0.189, relative to 0.130 for d' and 0.132 for P_r).

Consider model comparison evidence

As outlined at the beginning, knowing which measure is the more principled one to use in a given situation is a difficult task. Consideration of existing empirical ROC curves for the same or similar tasks will give a good indication of a reasonable measure (Pazzaglia, Dube, & Rotello, 2013; Rotello, Heit, & Dubé, 2015; Swets, 1986b). This suggestion is not as straightforward as it first appears given that ROCs derived from confidence rating procedures, in which participants provide a confidence judgement rather than a binary response, are unable to distinguish SDT and THT based accounts (Bröder, Kellen, Schütz, & Rohrmeier, 2013; Bröder & Schütz, 2009; Erdfelder & Buchner, 1998; Malmberg, 2002). The form of recognition memory ROCs derived from binary-response data, in which bias is manipulated via changing base rates or varying payoff, is also controversial (Bröder & Schütz, 2009; Dube & Rotello, 2012; Dube, Starns, Rotello, & Ratcliff, 2012; Kellen & Klauer, 2015), although recent work has suggested that the THT model gives a more parsimonious account of binary ROC recognition data (Kellen, Klauer, & Bröder, 2013).

Of course the decision, where possible, should not be solely based on the form of empirical ROCs and should take into account other sources of evidence. Recent work has moved beyond assessing the shape of ROCs to distinguish between models towards focusing on tests of critical predictions. For example, Province and Rouder (2012) focused on the prediction of conditional independence made by threshold models. Namely, that, while experimental manipulations may affect detection probabilities, once a particular state is obtained the distribution of responses is invariant. Province and Rouder (2012) used a confidence rating procedure to show that repeatedly presenting stimuli (words) affected the probability that an item was detected at test, but did not greatly affect the pattern of confidence ratings once a state had been entered (see also, Kellen & Klauer, 2015). Further, researchers are also beginning to incorporate reaction time data into their formal modelling to test crucial predictions (Dube et al., 2012; Province & Rouder, 2012). Reviewing this growing literature is beyond our scope here, but such evidence, where available, should factor into decision regarding choice of measure.

Finally, the models considered here reflect highly constrained versions of SDT and threshold accounts. We focused on these measures here as they are most often applied by researchers and practitioners alike. However, it is possible that we should re-evaluate our dependence on these measures and be ready to abandon them if background evidence suggests they may give a distorted pattern of means. This suggestion has been articulated elsewhere in reference to the common finding that z-ROCs have a slope less than 1. For example, Rotello et al. (2008) suggest the measures d_a and A_z as ones that take this asymmetry into account, however one draw back is that an estimate of the z-ROC slope is needed to compute these measures as this gives the ratio of the standard deviations of the old and new distributions.

In summary the best recommendation that can be made on the basis of the present work and extant literature is not a simple one; in order to state conclusions regarding experimental effects and interactions on recognition sensitivity the choice of measure must be thoroughly justified. If an adequate measure cannot be found on the basis of the existing knowledge, researchers may wish to analyse a range of

measures representing different theoretical accounts of the recognition process and acknowledge differences between measures. It is hoped that increasing knowledge of these issues will allow researchers to state their conclusions more clearly, even if the conclusions themselves are less clear.

Age-group by condition interactions

Returning to the primary motivation for the present simulations. There have been a number of working memory experiments reported in which the effect of age (i.e. the comparison of younger and older groups) was greater in conditions requiring the explicit binding of information relative to conditions with no such demands (Brown & Brockmole, 2010; Brown et al., 2016, Experiment 2; Peterson & Naveh-Benjamin, 2016; Isella et al., 2015). Given that this work uses the binary choice change detection task (or variants) the results of the present simulations cast great doubt on these findings and, in fact, imply that if the wrong measure is chosen these errors should be fairly commonplace. As reported in Chapter 3 a reanalysis of the data from Experiment 2 of Brown and Brockmole (2010) revealed that the effect size (η_P^2) estimate for the age $\times$ condition interaction was over 1.6 times larger with A' relative to proportion correct (THT) and disappeared altogether when analysed using a logit regression model. Further, although largely ignored by the authors, Isella et al. (2015) report an interaction with A' that is not present in their supplementary analysis of proportion correct.

Several studies have systematically manipulated sensitivity and bias with the single probe change detection task to produce ROC curves (Donkin et al., 2013, 2014; Rouder et al., 2008). These studies have found that the resulting ROCs conform nicely to the expectations of a THT model; they are linear with a unit slope. Recently, reaction time distributions in the standard change detection task have been shown to conform nicely to the expectations of discrete state models (Donkin et al., 2013). Thus, in the absence of more fine grained data, we may suggest that proportion correct or P_r are principled statistics for describing an observer's change detection performance uncontaminated by response bias. Thus the fact that

the evidence for the crucial interaction tends to arise with an unprincipled measure (A'), that can be criticised on multiple grounds, but is far less convincing with proportion correct should reassure us that there is no good evidence for a differential effect of age on the ability to bind features in VWM.

Of course it may be the case that, while the detection of feature changes is more consistent with a threshold process (Donkin et al., 2013; Rouder et al., 2008), detection of binding changes may draw upon more graded information. Thus an assessment of ROC curves for different variants of the change detection task is an important future step to ensure that measures of performance are as accurate as possible. This is not merely an academic exercise; as shown here, if sensitivity and response bias are confounded important differences in performance between groups (i.e. healthy and pathological ageing) may be obscured or groups may appear to differ when no true differences exist.

Chapter 9

General Discussion

The present thesis has focused on older adults' ability to integrate features in visual working memory and use these representations to detect subsequent changes to objects. In the process this work has revealed a number of methodological and measurement issues that have the potential to bias the assessment of this and related questions. Below the implications of our findings are discussed along with suggestions for further research.

9.1 Healthy Ageing and Binding in Working Memory

A primary aim of the current work was to assess potential boundary conditions under which healthy older adults may struggle to bind features and temporarily maintain the resulting representations in working memory. Several such boundary conditions presented themselves in the literature:

- Chapter 3 examined the role of presentation time in the emergence of a specific colour-shape binding deficit (cf. Brown & Brockmole, 2010). On the basis of previous theorising, it seemed possible that older adults are less able to make use of extra time to engage in effortful and elaborative processing of the conjunctions present in the study array (R. J. Allen et al., 2006; Mitchell, Johnson, Raye, Mather, & D'Esposito, 2000).

- Chapter 4 presented conjunction changes in blocks of trials alongside more salient changes to individual features. An increased reliance on familiarity based recognition would be expected to result in disproportionately poor sensitivity to binding changes in this condition relative to when these trials are presented separately (Cowan et al., 2006).

- Chapter 5 also assessed the role of mixing trial types but in addition used location as a to-be-remembered feature. Senescent change to the medial temporal lobes may be expected to produce a specific difficulty in remembering what was where (Mitchell, Johnson, Raye, & D'Esposito, 2000).

None of these potential boundary conditions were found to result in a specific age-related binding deficit. Analysis of raw accuracy revealed that the size of the binding cost (the difference between individual feature and binding accuracy) was not credibly different between younger and older adults; that is, specific contrasts invariably included zero within their HDIs. Analyses of sensitivity measures (P_r, d', k) with default Bayes Factors (Rouder et al., 2012) always preferred models omitting age $\times$ conditions interactions. Thus we were able to accumulate evidence *against* the existence of a specific age-related VWM deficit under the circumstances assessed, which had not been done before.

Of course, Bayes factors summarise the weight of evidence in favour of competing hypotheses conferred by the data. It is possible that there is a small age-related binding deficit but our relatively small sample sizes led us to favour the null hypothesis. Additional data may eventually lead to the accumulation of evidence for the interaction effect. To try and ease this concern we combined the data sets from Experiments 4, 6, and 7 (as the samples in Experiment 8 overlapped somewhat with Experiment 6) and assessed the relative evidence for the crucial interaction with 241 individuals (120 younger, 121 older). For both analyses of P_r and d' as outcome measures the winning model included main effects of condition (average of individual features versus binding) and age-group only whereas the second model contained the two-way interaction. Comparing these two models revealed approximately 10-to-1 evidence against the two-way interaction ($B_{1,2}$ = 9.67 and 12.534

for P_r and d', respectively). Taken as a whole, then, the present work offers strong evidence against the suggestion of a specific age-related binding deficit for simple features.

It is tempting to offer a number of speculative explanations as to why the present work found evidence against a specific VWM deficit for conjunctions, whereas previous work has reported evidence for this. In some cases these explanations are clearly necessary; for example in Chapter 5 we discussed the potential influence of cues in the probe arrays used by Cowan et al. (2006) that may have led to the appearance of a deficit that did not generalise to our more constrained single probe task. However, in many cases a major contribution of the present work is in putting the case that the weight of evidence in favour of VWM binding deficits was never particularly strong in the first place. A reanalysis of the second experiment reported by Brown and Brockmole (2010), with a more appropriate model for accuracy, provides no support for the interaction found with ANOVA (see Chapter 3) and in many cases the crucial evidence for interactions was either not sufficient (i.e. $p > 0.05$) or not provided (see Chapter 5). More worryingly the simulations presented in Chapter 8 imply that erroneous interactions should be rather commonplace unless the choice of outcome measure is fully justified, which it often is not. The broader measurement implications of this finding are outlined in more detail in the next section.

Nevertheless, there are further potential boundary conditions under which an age-related binding deficit may be observed. For example, Peich et al. (2013) found clear evidence for an increased frequency of mis-binding with age such that, when cued by location, older participants were more likely to recall a feature that was presented elsewhere. Accounting for age-differences in object memory it appears that older adults do not, in fact, commit more mis-bindings (Pertzov et al., 2015). Rather than reflecting a deficit in the formation and retention of integrated representations, the findings of Peich and colleagues may reflect some form of output interference as a result of participants having to cycle through multiple response options to reach the correct answer. A tendency to respond on the basis of familiarity (cf. M. G. Rhodes et al., 2008) may produce a specific age-effect on recall tasks such as these. As these

tasks are being developed for clinical application (Pertzov et al., 2013; Liang et al., 2016) the role of at-test interference in the sensitivity of this task to normal ageing should be a research priority.

As outlined in the opening chapter, location appears to occupy a privileged position in the processing of features and their integration (Ashby et al., 1996; Johnston & Pashler, 1990; Treisman & Gelade, 1980) and it has often been suggested that temporarily retaining what was where is a particular problem for older adults (R. J. Allen et al., 2013; Brockmole et al., 2008; Mitchell, Johnson, Raye, Mather, & D'Esposito, 2000). Experiments 7 and 8 of the present work, and the recent findings of Read et al. (2016), suggest that, at least for relatively simple features, older adults do not struggle to detect changes to feature-location conjunctions. Nevertheless, several forms of binding involving location have been discussed in the literature and it is possible that healthy ageing specifically affects some of these but not others. For example, allocentric spatial representation, that is detached from the viewpoint of the observer, depends on the hippocampus (O'Keefe & Nadel, 1978) but the extent to which egocentric, or viewpoint dependent, representation of space is dependent on this structure is unclear (see Baddeley, Jarrold, & Vargha-Khadem, 2011, for discussion). The tasks used here to assess VWM binding appear to promote the latter form of representation as items appear on a screen in front of the participant and changes involve a shift of an item relative to the observer. Of course, other levels of representation likely contribute to performance on this task, for example the categorical representation of the relative position of items (Postma et al., 2008), but it is possible to manipulate task demands to shift the contribution of egocentric and allocentric representation. For example, having participants mentally rotate item-location conjunctions during the retention interval would require a less viewpoint-dependent, and possibly more hippocampus dependent, representation.

If requiring participants to detach item-location bindings from their initially presented frame of reference leads to a specific age-effect then researchers may consider assessing this form of WM binding along with relational and conjunctive binding mechanisms. It has been suggested that binding to allocentric space is a relational

form of binding in which the object has to be related to positions in a cognitive map (Baddeley, Jarrold, & Vargha-Khadem, 2011). Consequently we may expect older adults that struggle to retain what was where in an allocentric WM task to also struggle to retain relations between items. This kind of simultaneous assessment of different kinds of association has the potential to resolve some of the discrepant findings in the literature and build upon our current understanding of the 'levels of binding', which are discussed in more detail below.

Firstly, potential further boundary conditions withstanding, how do we explain the present findings in reference to current theorising on feature binding in VWM and VWM more generally? Initially it had been thought that retaining feature conjunctions in VWM should require some additional resource above that required for the retention of individual features (Baddeley, 2000; Wheeler & Treisman, 2002). However, this has often been found not to be the case, as dual task requirements appear to disrupt binding change detection to the same extent as feature change detection (e.g. R. J. Allen et al., 2006, 2012; Delvenne et al., 2010; C. C. Morey & Bieler, 2013). This lowering of younger adults' change detection accuracy under dual task conditions appears to be a good simulation of the performance of our healthy older groups raising the possibility that reduced attentional resources are responsible for much of the age-difference in performance. However, the utility of vague explanatory concepts such as 'reduced resources' is questionable given their inherent flexibility and effort should be devoted to identifying additional constraints. Chapter 7 aimed to do just this by applying a simple processing model to our change detection data. Lapses of attention appear to make a significant contribution to age-differences in performance and, in addition, older adults appear to retain and use fewer items in VWM. As discussed in this chapter the cause of this reduced capacity is likely multifaceted and will take more fine grained investigation to pick apart.

Returning to feature binding specifically, according to the sampling theories of perception and VWM detailed in the first chapter, the combination of features into an integrated whole does not reflect the function of a separate process (as implied by Baddeley, 2000; Wheeler & Treisman, 2002, for example) but rather a limiting case

where attention is precise enough to localise features to the same object (Ashby et al., 1996; Vul & Rich, 2010). Indeed, Cowan et al. (2013) found that the estimated number of colour-shape conjunctions their participants could retain in VWM was approximated well by a simple account in which features were selected independently with their own capacity limits. This is in line with a number of findings demonstrating independent sampling of features in both perceptual (Bundesen et al., 2003; Kyllingsbæk & Bundesen, 2007; Vul & Rich, 2010) and VWM tasks (Bays et al., 2011; Fougnie, Asplund, & Marois, 2010; Fougnie, Cormiea, & Alvarez, 2013). For example, the error made in recalling the colour and orientation of an item following a brief delay appears to be uncorrelated (Bays et al., 2011), which would not be predicted if they were integral features, selected together (Bae & Flombaum, 2013; Fougnie & Alvarez, 2011; Garner, 1974).

An implication of this is that for an array of N items a participant who can retain k^{S} shapes and k^{C} colours is expected to sample $\frac{k^{\mathrm{S}}k^{\mathrm{C}}}{N}$ bindings. In this case the absence of a specific age-effect on feature binding found here is perhaps not surprising, as binding performance largely reflects the efficacy with which the individual features can be sampled and retained. However, it should be noted that the above equation *does* predict that a reduction in the number of features that can be retained will produce a disproportionate reduction in the number of sampled bindings. For example, a younger group who can retain, on average, 4.5 colours and 2.5 shapes (estimates from Chapter 7) would be expected to sample 1.88 bindings from 6 items, whereas an older group with 3.5 colours and 2 shapes in memory would get approximately 1.17 bindings. Thus a reduction of 22% and 20% in the number of colours and shapes that can be retained, respectively, is translated into a reduction of 38% for the number of bindings in VWM. This does not necessarily pose a problem if a principled processing model is applied to estimate the number of items in VWM from hits and false-alarms, however, this would affect patterns of accuracy. Indeed, simulations, not presented here, suggest that age by condition interactions can be produced with proportion correct as the outcome by merely reducing capacity for individual features. These effects seem rather small for reasonable age-effects on

feature capacity ($\sim$ 1 item), but may contribute to the occasional reporting of such a deficit.

Of course more work is required to more clearly establish to what extent the sampling of features is independent for multi-item arrays. Any correlation in the selection of features would invalidate these assumptions (Bundesen, 1990). One key, testable prediction that this independent sampling account makes, implied by the above equation, is that the effect of increasing set size should be disproportionately large for binding change detection accuracy. Another open question concerns how two features when sampled from the same object stay linked in VWM (Cowan et al., 2013). It is possible that shared location forms the basis of this but its influence appears to decrease over time (Logie et al., 2011), suggesting the need for another linking mechanism (for example the formation of 'object files' Kahneman, Treisman, & Gibbs, 1992).

If the independent sampling account does hold then the approach to understanding the feature binding deficit observed in early Alzheimer's disease will need rethinking. A speculative account can be offered. According to the theory of visual attention (TVA: Bundesen, 1990; Bundesen, Habekost, & Kyllingsbæk, 2011) the categorisation of objects as having a given feature proceeds in parallel as a race between the possible categorisations (i.e. feature dimensions) with the rate with which a particular characterisation is made influenced by the decision weight allocated to (or bias towards) certain objects or feature dimensions in the array. In our tasks we may assume that participants do not have a particular, systematic bias towards certain objects in the display (given random placement of items) or certain features of those objects (given both are needed to detect a feature swap). Nevertheless, these object and feature biases are presumably under attentional control. If certain groups exhibit a specific bias towards a particular feature dimension when selecting from multi-item displays, binding change detection will suffer. There is some evidence from the application of the TVA, that participants with mild cognitive impairment and mild-AD exhibit greater, or more volatile, bias in attentional selection (K. Finke, Myers, Bublak, & Sorg, 2013; Redel et al., 2012). How these findings

relate to their VWM binding deficit is a fascinating open question. Future work may consider assessing a change detection task in which *either* feature dimension may change (i.e. that introduced by Luck & Vogel, 1997). If patients exhibit a deficit here it may suggest that it is not a deficit of binding per se but rather a deficit of parallel feature selection from multi-item displays.

Healthy ageing and levels of binding

Chapter 1 summarises the wealth of research conducted on the associative deficit observed in normal ageing. Retaining the relation of distinct items reflects a different theoretical 'level of binding' from the retention of object feature conjunctions (Zimmer & Ecker, 2010; Zimmer et al., 2006). The former is said to rely on extrinsic binding between an item and its surrounding context whereas the latter level serves to bind the intrinsic, defining features of objects. Chapter 6 aimed to directly compare these two levels while avoiding the common confound between the type of binding required and the complexity of the stimuli (cf. T. Chen & Naveh-Benjamin, 2012). The findings of this preliminary experiment were not as clear as one may have hoped. There was some suggestion that age-differences in performance (as measured by area under the ROC curve) were greater for pairings of colour and shape when these features were presented in distinct 'items' as opposed to when they were presented within the same 'item'. However, the driver of this tendency appeared to go against the common expression of the associative deficit. Older adults were less able to recognise intact pairings in the extrinsic condition, whereas the associative deficit is seen, in both LTM and VWM, as a tendency to falsely recognise recombined lures (e.g., Bender et al., 2010; Castel & Craik, 2003; T. Chen & Naveh-Benjamin, 2012; M. G. Rhodes et al., 2008). Additional data on this question are clearly needed and the paradigm developed for this experiment should provide a good starting point.

If future studies build on the preliminary findings presented here and show that extrinsic binding *is* specifically affected by healthy ageing then researchers may consider assessing how this relates to measures of recollection and familiarity. These two distinct phenomenological sensations that accompany remembering share a the-

oretical link to the different hypothetical levels of binding (Zimmer & Ecker, 2010). Further these memorial experiences appear to dissociate in healthy ageing with a pronounced age effect on recollection and little-to-no effect of age on familiarity (Koen & Yonelinas, 2014; McCabe et al., 2009). This is clear in older adults' increased tendency towards responding to recognition probes on the basis of feelings of knowing (Mäntylä, 1993). Whether or not the apparent preservation of intrinsic binding mechanisms, as demonstrated here, relates to this recollection-familiarity shift is an important question. Assessing performance on a variety of measure of binding and recognition has the potential to unify these distinct findings in the cognitive ageing literature.

The relation between levels of binding and the distinction of recollection and familiarity also has potential practical implications. Like the assessment of short-term memory binding (e.g. Parra, Abrahams, Logie, & Della Sala, 2010), tasks probing different forms of recognition appear to distinguish between healthy older adults and mild stages of AD (Algarabel et al., 2009; Tse et al., 2010). Promisingly, a recent meta analysis found that, with an objective measure of recollection of familiarity (such as inclusion/ exclusion tasks) as opposed to subjective reports of remembering and knowing, healthy older adults exhibit preserved familiarity with impaired recollection, whereas those with mild to moderate AD exhibit impairments in *both* of these dimensions (Koen & Yonelinas, 2014). Thus it is possible that conducting simultaneous assessments of temporary shape-colour binding and the reliance on familiarity-based recognition will make for a more sensitive neuropsychological assessments for distinguishing healthy and pathological ageing.

9.2 Measurement Implications

In assessing the effect of healthy ageing on the ability to retain feature conjunctions we have encountered a couple of measurement issues that may affect one's ability to address this question. The first was methodological, and concerned the nature of the test probe in change detection tasks, whereas the second was statistical, and affects the interpretation of interaction effects across studies using different outcome

measures.

Previous work has suggested that the manner in which VWM is probed can have a fairly dramatic effect on patterns of performance (Alvarez & Thompson, 2009; Makovski et al., 2010). Wheeler and Treisman (2002) also found this, as they observed that proportion correct was roughly equivalent between shape-only and binding conditions when VWM was probed with a single test item, but binding performance was relatively low when probed with a whole display (see also, Kondo & Saiki, 2012; Yeh et al., 2005). They proposed that feature bindings were specifically susceptible to the distracting effects of having to process multiple items at test, whereas others have proposed that the whole display test may support VWM for features (J. S. Johnson et al., 2008). However, it is difficult to directly compare different probing methods using proportion correct given the inherent constraints these methods place on the use of information from VWM (Cowan et al., 2013; H. Zhang et al., 2010). Using simple processing models we found no evidence for binding specific interference in the whole display task (Experiments 1 and 2). Rather, when we manipulated the tasks themselves in order to better match their demands (Experiment 3), we found no evidence that the number of test items affected accuracy. This is in line with other studies showing that VWM is robust to different methods of testing (Woodman et al., 2012).

In developing the processing models for the whole display task, and for the exploratory modelling in Chapter 7, we encountered a number of open questions. Firstly, it is not clear whether participants can make use of partial information in the whole display task. Using the single probe task, Rouder et al. (2008) varied set size and change probability to trace out isosensitivity and isobias curves and found them to be straight, as predicted by the discrete state model with no partial information (see also Donkin et al., 2013, 2014). To our knowledge, no such assessment has been made of the whole display paradigm, although Wilken and Ma (2004) used a confidence rating procedure, rather than manipulating expectation of a change, to draw out their ROC curves which cannot adequately discriminate between high-threshold and detection theory accounts (see, e.g., Malmberg, 2002). It is possible

that participants make use of partial information with a whole display, for example knowledge of the ensemble statistics of an array (Brady & Alvarez, 2011). If this is the case, assessment of interactions with discrete-state measures, such as proportion correct or P_r, may lead to erroneous conclusions (as shown in Chapter 8 of this thesis). For this reason, and to maintain consistency with previous assessments of age-differences in VWM, we opted for the single probe task.

Another open question concerns the nature of guessing in change detection. Assuming whole display performance is mediated by discrete states, it is possible that observers use knowledge of the number of items in VWM (k) and the number of items presented (N) to guide response selection in an uncertain state. As we outline in Chapter 7 this could also be true for the single probe task introduced by Wheeler and Treisman (2002). Our preliminary model comparison evidence suggests that this may be the case. Accurate measurement of the number of items observers can retain in VWM requires proper characterisation of the observer's guessing strategy, otherwise estimates can be distorted (see Hardman & Cowan, 2016). Studies manipulating set size along with probability of a change, or the number of to-be-detected changes, are needed to distinguish between informed and uninformed guessing in change detection.

More profound measurement implications come from the simulations presented in Chapter 8. Previous work had demonstrated that, in the presence of differences in response bias, analysis of an unprincipled measure of sensitivity leads to erroneous conclusions regarding main-effects (Rotello et al., 2008; Schooler & Shiffrin, 2005). Our simulations show that, even in the absence of variation in response bias, analysis of an inappropriate measure can lead to shockingly high type I error rates for tests of interactions, provided that both factors produce moderate-to-large main effects. Additional variation in response bias between groups and conditions confounds this pattern further. Choice of A' in particular is problematic as, despite its regularly claimed non-parametric foundations, it implies specific evidence distributions (Macmillan & Creelman, 1996; Pastore et al., 2003) and does not measure the ROC area as once claimed (Smith, 1995; J. Zhang & Mueller, 2005). That so

many specific age-related deficits of VWM feature binding have been reported with A' as the primary outcome measure should be cause for concern (e.g. Brown & Brockmole, 2010; Isella et al., 2015; Peterson & Naveh-Benjamin, 2016).

These findings are a cause for concern more broadly. Tests of interactions are a key inferential tool in cognitive psychology and can occasionally be interpreted unambiguously. For example, cross over interactions—that is, the opposite pattern of effects in one group/ condition relative to another (such as interactions in some dual tasking studies, e.g., Darling, Della Sala, & Logie, 2009)—cannot be transformed away by scaling the outcome variable. This is not true for interactions where an effect is magnified or diminished in a certain group or condition. Unfortunately these additive interactions are by far the most common in cognitive ageing and neuropsychological research (Salthouse, 2000). Typically one group that performs lower overall is found to perform specifically poorly under certain conditions. Such dissociations can then inform cognitive theory. The findings of our simulations underline the fact that such interactions can easily be produced or erased by change of measurement scale (Salthouse, 2000).

For researchers using recognition or detection paradigms, consideration of ROC curves and other model comparison evidence will help guide selection of a principled measure. Alternatively, adding a simple confidence rating to the trial sequence allows researchers to trace out an empirical ROC curve and conduct a truly non-parametric assessment of sensitivity, as we did in Chapter 6. These methodological considerations will lead to better assessment of differences between groups and will prevent the propagation of errors throughout the literature.

Conclusion Healthy ageing affects a number of cognitive processes and in particular memory for inter-item associations which support the feeling of episodic remembering. Retaining conjunctions of simple features over a brief period in order to perform a recognition judgement appears to be largely invariant to the effects of healthy ageing, relative to the more general effect on temporary visual storage (i.e. features alone). The present work strengthens the suggestion that feature binding in VWM is age-invariant, both by critically evaluating the current literature and

providing new data bearing on this question. While future work may still establish boundary conditions on this invariance, the collective failure to find a specific deficit under a wide range of circumstances strengthens the suggestion that a feature binding deficit is a specific hallmark of early Alzheimer's disease. Future work bearing on this and similar questions should be wary of measurement assumptions when assessing group-differences in recognition or detection performance. Failure to properly justify an outcome measure may lead to the conclusion of a specific age-related deficit when, in fact, none exists.

References

Aaronson, D., & Watts, B. (1987). Extensions of Grier's computational formulas for A' and B" to below-chance performance. *Psychological Bulletin*, *102*(3), 439–442.

Adam, K. C., Mance, I., Fukuda, K., & Vogel, E. K. (2015). The contribution of attentional lapses to individual differences in visual working memory capacity. *Journal of Cognitive Neuroscience*, *27*(8), 1601–1616.

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Algarabel, S., Escudero, J., Mazón, J. F., Pitarque, A., Fuentes, M., Peset, V., & Lacruz, L. (2009). Familiarity-based recognition in the young, healthy elderly, mild cognitive impaired and Alzheimer's patients. *Neuropsychologia*, *47*(10), 2056–2064.

Allen, P. A., Sliwinski, M., Bowie, T., & Madden, D. J. (2002). Differential age effects in semantic and episodic memory. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *57*(2), P173–P186.

Allen, R. J., Baddeley, A. D., & Hitch, G. J. (2006). Is the binding of visual features in working memory resource-demanding? *Journal of Experimental Psychology: General*, *135*(2), 298–313.

Allen, R. J., Brown, L. A., & Niven, E. (2013). Aging and visual feature binding in working memory. In H. St. Clair-Thompson (Ed.), *Working memory: Developmental differences, component processes and improvement mechanisms* (pp. 83–96). New York, NY: Nova Science Publishers.

Allen, R. J., Castellà, J., Ueno, T., Hitch, G. J., & Baddeley, A. D. (2015). What does visual suffix interference tell us about spatial location in working memory? *Memory & Cognition*, *43*(1), 133–142.

Allen, R. J., Hitch, G. J., Mate, J., & Baddeley, A. D. (2012). Feature binding and attention in working memory: A resolution of previous contradictory findings. *The Quarterly Journal of Experimental Psychology*, *65*(12), 2369–2383.

Allen, R. J., Vargha-Khadem, F., & Baddeley, A. D. (2014). Item-location binding

in working memory: Is it hippocampus-dependent? *Neuropsychologia*, *59*, 74–84.

Alvarez, G. A., & Thompson, T. W. (2009). Overwriting and rebinding: Why feature-switch detection tasks underestimate the binding capacity of visual working memory. *Visual Cognition*, *17*(1-2), 141–159.

Arnold, P. G., & Bower, G. H. (1972). Perceptual conditions affecting ease of association. *Journal of experimental psychology*, *93*(1), 176–180.

Asch, S. E., Ceraso, J., & Heimer, W. (1960). Perceptual conditions of association. *Psychological Monographs: General and Applied*, *74*(3), 1–48.

Ashby, F. G., Prinzmetal, W., Ivry, R., & Maddox, W. T. (1996). A formal theory of feature binding in object perception. *Psychological Review*, *103*(1), 165–192.

Baddeley, A. D. (1982). Domains of recollection. *Psychological Review*, *89*(6), 708–729.

Baddeley, A. D. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423.

Baddeley, A. D. (2007). *Working memory, thought, and action*. Oxford University Press.

Baddeley, A. D., Allen, R., & Vargha-Khadem, F. (2010). Is the hippocampus necessary for visual and verbal binding in working memory? *Neuropsychologia*, *48*, 1089–1095.

Baddeley, A. D., Allen, R. J., & Hitch, G. J. (2011). Binding in visual working memory: The role of the episodic buffer. *Neuropsychologia*, *49*(6), 1393–1400.

Baddeley, A. D., Jarrold, C., & Vargha-Khadem, F. (2011). Working memory and the hippocampus. *Journal of Cognitive Neuroscience*, *23*(12), 3855–3861.

Badham, S. P., Estes, Z., & Maylor, E. A. (2012). Integrative and semantic relations equally alleviate age-related associative memory deficits. *Psychology and Aging*, *27*(1), 141–152.

Bae, G. Y., & Flombaum, J. I. (2013). Two items remembered as precisely as one: How integral features can improve visual working memory. *Psychological Science*, *24*(10), 2038–2047.

Bastin, C., Bahri, M. A., Miévis, F., Lemaire, C., Collette, F., Genon, S., ... Salmon, E. (2014). Associative memory and its cerebral correlates in Alzheimer's disease: Evidence for distinct deficits of relational and conjunctive memory. *Neuropsychologia*, *63*, 99–106.

Bastin, C., Diana, R. A., Simon, J., Collette, F., Yonelinas, A. P., & Salmon, E. (2013). Associative memory in aging: the effect of unitization on source memory. *Psychology and Aging*, *28*(1), 275–283.

Bastin, C., & Van der Linden, M. (2006). The effects of aging on the recognition of different types of associations. *Experimental Aging Research*, *32*(1), 61–77.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4 [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=lme4` (R package version 1.1-7)

Bayen, U. J., Phelps, M. P., & Spaniol, J. (2000). Age-related differences in the use of contextual information in recognition memory a global matching approach. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *55*(3), P131–P141.

Bays, P. M., Catalao, R. F., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, *9*(10), 7.

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science*, *851*, 851–854.

Bays, P. M., Wu, E. Y., & Husain, M. (2011). Storage and binding of object features in visual working memory. *Neuropsychologia*, *49*, 1622–1631.

Bender, A. R., Naveh-Benjamin, M., & Raz, N. (2010). Associative deficit in recognition memory in a lifespan sample of healthy adults. *Psychology and Aging*, *25*(4), 940–948.

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of P values and evidence. *Journal of the American Statistical Association*, *82*(397), 112–122.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens,

M. H. H., & White, J.-S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, *24*(3), 127–135.

Bopp, K. L., & Verhaeghen, P. (2005). Aging and verbal memory span: A meta-analysis. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *60*(5), P223–P233.

Bopp, K. L., & Verhaeghen, P. (2009). Working memory and aging: separating the effects of content and context. *Psychology and Aging*, *24*(4), 968–980.

Borg, C., Leroy, N., Favre, E., Laurent, B., & Thomas-Antérion, C. (2011). How emotional pictures influence visuospatial binding in short-term memory in ageing and Alzheimer's disease? *Brain and Cognition*, *76*(1), 20–25.

Bowles, R. P., & Salthouse, T. A. (2003). Assessing the age-related effects of proactive interference on working memory tasks using the rasch model. *Psychology and Aging*, *18*(3), 608–615.

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science*, *22*(3), 384–392.

Brockmole, J. R., & Logie, R. H. (2013). Age-related change in visual working memory: a study of 55,753 participants aged 8–75. *Frontiers in Psychology*, *4*, 12. doi: 10.3389/fpsyg.2013.00012

Brockmole, J. R., Parra, M. A., Della Sala, S., & Logie, R. H. (2008). Do binding deficits account for age-related decline in visual working memory? *Psychonomic Bulletin and Review*, *15*(3), 543–547.

Bröder, A., Kellen, D., Schütz, J., & Rohrmeier, C. (2013). Validating a two-high-threshold measurement model for confidence rating data in recognition. *Memory*, *21*(8), 916–944.

Bröder, A., & Schütz, J. (2009). Recognition ROCs are curvilinear—or are they? on premature arguments against the two-high-threshold model of recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 587–606.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455.

Brown, L. A., & Brockmole, J. R. (2010). The role of attention in binding visual features in working memory: Evidence from cognitive ageing. *The Quaterly Journal of Experimental Psychology*, *63*(10), 2067–2079.

Brown, L. A., Niven, E., Logie, R. H., Rhodes, S., & Allen, R. J. (2016). Visual feature binding in younger and older adults: Encoding and suffix interference effects. *Memory*.

Buckner, R. L., Snyder, A. Z., Shannon, B. J., LaRossa, G., Sachs, R., Fotenos, A. F., ... others (2005). Molecular, structural, and functional characterization of Alzheimer's disease: evidence for a relationship between default activity, amyloid, and memory. *The Journal of Neuroscience*, *25*(34), 7709–7717.

Bundesen, C. (1990). A theory of visual attention. *Psychological Review*, *97*(4), 523–547.

Bundesen, C., Habekost, T., & Kyllingsbæk, S. (2011). A neural theory of visual attention and short-term memory (NTVA). *Neuropsychologia*, *49*(6), 1446–1457.

Bundesen, C., Kyllingsbæk, S., & Larsen, A. (2003). Independent encoding of colors and shapes from two stimuli. *Psychonomic Bulletin & Review*, *10*(2), 474–479.

Burgess, N., Maguire, E. A., & O'Keefe, J. (2002). The human hippocampus and spatial and episodic memory. *Neuron*, *35*(4), 625–641.

Cant, J. S., & Goodale, M. A. (2007). Attention to form or surface properties modulates different regions of human occipitotemporal cortex. *Cerebral Cortex*, *17*(3), 713–731.

Cant, J. S., Large, M.-E., McCall, L., & Goodale, M. A. (2008). Independent processing of form, colour, and texture in object perception. *Perception*, *37*, 57–78.

Carriere, J. S., Cheyne, J. A., Solman, G. J., & Smilek, D. (2010). Age trends for

failures of sustained attention. *Psychology and Aging*, *25*(3), 569–574.

Castel, A. D., & Craik, F. I. M. (2003). The effects of aging and divided attention on memory for item and associative information. *Psychology and Aging*, *18*(4), 873–885.

Ceraso, J., Kourtzi, Z., & Ray, S. (1998). The integration of object properties. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(5), 1152–1161.

Chalfonte, B. L., & Johnson, M. K. (1996). Feature memory and binding in young and older adults. *Memory and Cognition*, *24*(4), 403–416.

Chen, T., & Naveh-Benjamin, M. (2012). Assessing the associative deficit of older adults in long-term and short-term/working memory. *Psychology and Aging*, *27*(3), 666–682.

Chen, Z., & Cowan, N. (2013). Working memory inefficiency: Minimal information is utilized in visual recognition tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(5), 1449–1462.

Cohen, A., & Ivry, R. (1989). Illusory conjunctions inside and outside the focus of attention. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(4), 650–663.

Cohen, A., & Ivry, R. B. (1991). Density effects in conjunction search: evidence for a coarse location mechanism of feature integration. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(4), 891–901.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87–185.

Cowan, N., Blume, C. L., & Saults, J. S. (2013). Attention to attributes and objects in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 731–747.

Cowan, N., Elliott, E. M., Saults, J. S., Morey, C. C., Mattox, S., Hismjatullina, A., & Conway, A. R. A. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology*, *51*(1), 42–100.

Cowan, N., Hardman, K., Saults, J. S., Blume, C. L., Clark, K. M., & Sunday, M. A. (2016). Detection of the number of changes in a display in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(2), 169–185.

Cowan, N., Naveh-Benjamin, M., Kilb, A., & Saults, J. S. (2006). Life-span development of visual working-memory: When is feature binding difficult? *Developmental Psychology*, *42*(6), 1089–1102.

Cowan, N., & Rouder, J. N. (2009). Comment on "dynamic shifts of limited working memory resources in human vision". *Science*, *323*(5916), 877c.

Cowan, N., Saults, J. S., & Blume, C. L. (2014). Central and peripheral components of working memory storage. *Journal of Experimental Psychology: General*, *143*(5), 1806–1836.

Craik, F. I. M. (2006). Remembering items and their contexts: effects of aging and divided attention. In H. D. Zimmer, A. Mecklinger, & U. Lindenberger (Eds.), *Handbook of binding and memory: perspectives from cognitive neuroscience* (pp. 571–594). New York, NY: Oxford University Press.

Craik, F. I. M., & Bialystok, E. (2006). Cognition through the lifespan: Mechanisms of change. *Trends in Cognitive Sciences*, *10*(3), 131–138.

Craik, F. I. M., & Byrd, M. (1982). Aging and cognitive deficits: The role of attentional resources. In F. I. M. Craik & S. Trehub (Eds.), *Aging and cognitive processes* (pp. 191–211). New York: Plenum.

Craik, F. I. M., Luo, L., & Sakuta, Y. (2010). Effects of aging and divided attention on memory for items and their contexts. *Psychology and Aging*, *25*(4), 968–979.

Craik, F. I. M., & Rabinowitz, J. C. (1985). The effects of presentation rate and encoding task on age-related memory deficits. *Journal of Gerontology*, *40*(3), 309–315.

Cramer, A. O., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R. P., ... Wagenmakers, E.-J. (2016). Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies. *Psychonomic Bulletin*

& Review, *23*(2), 640–647.

Cumming, G. (2013). The new statistics: Why and how. *Psychological Science*, *12*, 7–29.

Cumming, G., & Finch, S. (2005). Inference by eye: confidence intervals and how to read pictures of data. *American Psychologist*, *60*(2), 170–180.

Darling, S., Della Sala, S., & Logie, R. H. (2009). Dissociation between appearance and location within visuo-spatial working memory. *The Quarterly Journal of Experimental Psychology*, *62*(3), 417–425.

Das, S. R., Mancuso, L., Olson, I. R., Arnold, S. E., & Wolk, D. A. (2015). Short-term memory depends on dissociable medial temporal lobe regions in amnestic mild cognitive impairment. *Cerebral Cortex*, online ahead of print.

Davachi, L. (2006). Item, context and relational episodic encoding in humans. *Current Opinion in Neurobiology*, *16*(6), 693–700.

Della Sala, S., Parra, M. A., Fabi, K., Luzzi, S., & Abrahams, S. (2012). Short-term memory binding is impaired in AD but not in non-AD dementias. *Neuropsychologia*, *50*, 833–840.

Delvenne, J. F., & Bruyer, R. (2004). Does visual short-term memory store bound features? *Visual Cognition*, *11*(1), 1–27.

Delvenne, J. F., Cleeremans, A., & Laloyaux, C. (2010). Feature bindings are maintained in visual short-term memory without sustained focused attention. *Experimental Psychology*, *57*(2), 108–116.

Diana, R. A., Yonelinas, A. P., & Ranganath, C. (2007). Imaging recollection and familiarity in the medial temporal lobe: a three-component model. *Trends in Cognitive Sciences*, *11*(9), 379–386.

Didic, M., Barbeau, E. J., Felician, O., Tramoni, E., Guedj, E., Poncet, M., & Ceccaldi, M. (2011). Which memory system is impaired first in Alzheimer's disease? *Journal of Alzheimer's Disease*, *27*(1), 11–22.

Di Lollo, V. (2012). The feature-binding problem is an ill-posed problem. *Trends in Cognitive Sciences*, *16*(6), 317–321.

Dixon, P. (2008). Models of accuracy in repeated-measures designs. *Journal of*

Memory and Language, *59*(4), 447–456.

Donkin, C., Nosofsky, R. M., Gold, J. M., & Shiffrin, R. M. (2013). Discrete-slots models of visual working-memory response times. *Psychological Review*, *120*(4), 873–902.

Donkin, C., Tran, S. C., & Nosofsky, R. (2014). Landscaping analyses of the ROC predictions of discrete-slots and signal-detection models of visual working memory. *Attention, Perception, & Psychophysics*, *76*(7), 2103–2116.

Dube, C., & Rotello, C. M. (2012). Binary ROCs in perception and recognition memory are curved. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(1), 130–151.

Dube, C., Starns, J. J., Rotello, C. M., & Ratcliff, R. (2012). Beyond ROC curvature: Strength effects and response time data support continuous-evidence models of recognition memory. *Journal of Memory and language*, *67*(3), 389–406.

Duncan, J. (1984). Selective attention and the organization of visual information. *Journal of Experimental Psychology: General*, *113*(4), 501–517.

Dvorine, I. (1963). Quantitative classification of color blind. *Journal of General Psychology*, *68*, 255–265.

Ecker, U. K., Maybery, M., & Zimmer, H. D. (2013). Binding of intrinsic and extrinsic features in working memory. *Journal of Experimental Psychology: General*, *142*(1), 218–234.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*(3), 193–242.

Eichenbaum, H., Yonelinas, A., & Ranganath, C. (2007). The medial temporal lobe and recognition memory. *Annual Review of Neuroscience*, *30*, 123–152.

Ekstrom, A. D., Kahana, M. J., Caplan, J. B., Fields, T. A., Isham, E. A., Newman, E. L., & Fried, I. (2003). Cellular networks underlying human spatial navigation. *Nature*, *425*(6954), 184–188.

Elsley, J. V., & Parmentier, F. B. (2009). Is verbal-spatial binding in working memory impaired by a concurrent memory load? *The Quarterly Journal of Experimental Psychology*, *62*(9), 1696–1705.

Emery, L., Hale, S., & Myerson, J. (2008). Age differences in proactive interference, working memory, and abstract reasoning. *Psychology and Aging*, *23*(3), 634–645.

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models: A review of the literature. *Zeitschrift für Psychologie/Journal of Psychology*, *217*(3), 108–124.

Erdfelder, E., & Buchner, A. (1998). Process-dissociation measurement models: Threshold theory or detection theory? *Journal of Experimental Psychology: General*, *127*(1), 83–96.

Fandakova, Y., Sander, M. C., Werkle-Bergner, M., & Shing, Y. L. (2014). Age differences in short-term memory binding are related to working memory performance across the lifespan. *Psychology and Aging*, *29*(1), 140–149.

Fiacconi, C. M., & Milliken, B. (2013). Visual memory for feature bindings: The disruptive effect of responding to new perceptual input. *The Quarterly Journal of Experimental Psychology*, *66*(8), 1572–1600.

Finke, C., Braun, M., Ostendorf, F., Lehmann, T.-N., Hoffmann, K.-T., Kopp, U., & Ploner, C. J. (2008). The human hippocampal formation mediates short-term memory of colour–location associations. *Neuropsychologia*, *46*(2), 614–623.

Finke, K., Myers, N., Bublak, P., & Sorg, C. (2013). A biased competition account of attention and memory in Alzheimer's disease. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *368*(1628), 20130062.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). Mini-mental state: A practical method for grading the cognitive state of patients for the clinician. *Journal Psychiatric Research*, *12*, 189–198.

Fougnie, D., & Alvarez, G. A. (2011). Object features fail independently in visual working memory: Evidence for a probabilistic feature-store model. *Journal of Vision*, *11*(12), 3. doi: 10.1167/11.12.3

Fougnie, D., Asplund, C. L., & Marois, R. (2010). What are the units of storage in visual working memory? *Journal of Vision*, *10*(12), 27. doi: 10.1167/10.12.27

Fougnie, D., Cormiea, S. M., & Alvarez, G. A. (2013). Object-based benefits without

object-based representations. *Journal of Experimental Psychology: General*, *142*(3), 621–626.

Fougnie, D., & Marois, R. (2009). Attentive tracking disrupts feature binding in visual working memory. *Visual Cognition*, *17*(1-2), 48–66.

Frank, D. J., Nara, B., Zavagnin, M., Touron, D. R., & Kane, M. J. (2015). Validating older adults' reports of less mind-wandering: An examination of eye movements and dispositional influences. *Psychology and Aging*, *30*(2), 266–278.

Gajewski, D. A., & Brockmole, J. R. (2006). Feature bindings endure without attention: Evidence from an explicit recall task. *Psychonomic Bulletin & Review*, *13*(4), 581–587.

Garner, W. R. (1974). *The processing of information and structure*. Potomac, MD: Lawrence Erlbaum Associates.

Gazzaley, A., Clapp, W., Kelley, J., McEvoy, K., Knight, R. T., & D'Esposito, M. (2008). Age-related top-down suppression deficit in the early stages of cortical visual memory processing. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(35), 13122–13126.

Gazzaley, A., Cooney, J. W., Rissman, J., & D'Esposito, M. (2005). Top-down suppression deficit underlies working memory impairment in normal aging. *Nature Neuroscience*, *8*(10), 1298–1300.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and hierarchical/multilevel models*. Cambridge, UK: Cambridge University Press.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 1360–1383.

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, *60*(4), 328–331.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606.

Gilchrist, A. L., & Cowan, N. (2014). A two-stage search of visual working memory: investigating speed in the change-detection paradigm. *Attention, Perception, & Psychophysics*, *76*(7), 2031–2050.

Glisky, E. L. (2007). Changes in cognitive function in human aging. In D. L. Riddle (Ed.), *Brain aging: Models, methods, and mechanisms* (pp. 3–20). Boca Raton, Florida, USA: CRC Press.

Green, D. M. (1964). General prediction relating yes-no and forced-choice results [abstract]. *The Journal of the Acoustical Society of America*, *36*(5), 1042–1042.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Grier, J. B. (1971). Nonparametric indexes for sensitivity and bias: computing formulas. *Psychological Bulletin*, *75*(6), 424–429.

Griffin, I. C., & Nobre, A. C. (2003). Orienting attention to locations in internal representations. *Journal of Cognitive Neuroscience*, *15*(8), 1176–1194.

Hardman, K. O., & Cowan, N. (2015). Remembering complex objects in visual working memory: do capacity limits restrict objects or features? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(2), 325–347.

Hardman, K. O., & Cowan, N. (2016). Reasoning and memory: People make varied use of the information available in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, in press.

Hartman, M., & Warren, L. H. (2005). Explaining age differences in temporal working memory. *Psychology and Aging*, *20*(4), 645–656.

Hartshorne, J. K. (2008). Visual working memory capacity and proactive interference. *PLoS one*, *3*(7), e2716.

Hasher, L., & Zacks, R. T. (1988). Working memory, comprehension, and aging: A review and a new view. *Psychology of Learning and Motivation*, *22*, 193–225.

Haxby, J. V., Petit, L., Ungerleider, L. G., & Courtney, S. M. (2000). Distinguishing the functional roles of multiple regions in distributed neural systems for visual working memory. *Neuroimage*, *11*(2), 145–156.

Healey, M. K., & Kahana, M. J. (2016). A four-component model of age-related memory change. *Psychological Review*, *123*(1), 23–69.

Heywood, C. A., & Kentridge, R. W. (2003). Achromatopsia, color vision, and cortex. *Neurologic Clinics*, *21*(2), 483–500.

Hodos, W. (1970). Nonparametric index of response bias for use in detection and recognition experiments. *Psychological Bulletin*, *74*(5), 351–354.

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*(5), 1157–1164.

Hu, Y., Hitch, G. J., Baddeley, A. D., Zhang, M., & Allen, R. J. (2014). Executive and perceptual attention play different roles in visual working memory: Evidence from suffix and strategy effects. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(4), 1665–1678.

Hubel, D. H., & Livingstone, M. S. (1987). Segregation of form, color, and stereopsis in primate area 18. *The Journal of Neuroscience*, *7*(11), 3378–3415.

Humphrey, G. K., Goodale, M. A., Jakobson, L. S., & Servos, P. (1994). The role of surface information in object recognition: Studies of a visual form agnosic and normal subjects. *Perception*, *23*, 1457–1457.

Hyun, J.-S., Woodman, G. F., Vogel, E. K., Hollingworth, A., & Luck, S. J. (2009). The comparison of visual working memory representations with perceptual inputs. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(4), 1140–1160.

Isella, V., Molteni, F., Mapelli, C., & Ferrarese, C. (2015). Short term memory for single surface features and bindings in ageing: A replication study. *Brain and Cognition*, *96*, 38–42.

Jackson, J. D., & Balota, D. A. (2012). Mind-wandering in younger and older adults: converging evidence from the sustained attention to response task and reading for comprehension. *Psychology and Aging*, *27*(1), 106–119.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and*

Language, *59*(4), 434–446.

Jeneson, A., Mauldin, K. N., & Squire, L. R. (2010). Intact working memory for relational information after medial temporal lobe damage. *The Journal of Neuroscience*, *30*(41), 13624–13629.

Jenkins, L., Myerson, J., Joerding, J. A., & Hale, S. (2000). Converging evidence that visuospatial cognition is more age-sensitive than verbal cognition. *Psychology and Aging*, *15*(1), 157–175.

Jennings, J. M., & Jacoby, L. L. (1993). Automatic versus intentional uses of memory: aging, attention, and control. *Psychology and Aging*, *8*(2), 283–293.

Jiang, Y., Olson, I. R., & Chun, M. M. (2000). Organization of visual short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(3), 683–702.

Johnson, J. S., Hollingworth, A., & Luck, S. J. (2008). The role of attention in the maintenance of feature bindings in visual short-term memory. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(1), 41–55.

Johnson, M. K. (1992). MEM: mechanisms of recollection. *Journal of Cognitive Neuroscience*, *4*(3), 268–280.

Johnson, M. K., McMahon, R. P., Robinson, B. M., Harvey, A. N., Hahn, B., Leonard, C. J., ... Gold, J. M. (2013). The relationship between working memory capacity and broad measures of cognitive ability in healthy adults and people with schizophrenia. *Neuropsychology*, *27*(2), 220–229.

Johnson, W., Logie, R. H., & Brockmole, J. R. (2010). Working memory tasks differ in factor structure across age cohorts: Implications for dedifferentiation. *Intelligence*, *38*, 513–528.

Johnston, J. C., & Pashler, H. (1990). Close binding of identity and location in visual feature perception. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(4), 843–856.

Jost, K., Bryck, R. L., Vogel, E. K., & Mayr, U. (2011). Are old adults just like low working memory young adults? Filtering efficiency and age differences in visual working memory. *Cerebral Cortex*, *21*, 1147–1154.

Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology*, *24*, 175–219.

Kass, R. E., Carlin, B. P., Gelman, A., & Neal, R. M. (1998). Markov chain Monte Carlo in practice: a roundtable discussion. *The American Statistician*, *52*(2), 93–100.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Kausler, D. H., & Puckett, J. M. (1981). Adult age differences in memory for sex of voice. *Journal of Gerontology*, *36*(1), 44–50.

Kellen, D., & Klauer, K. C. (2015). Signal detection and threshold modeling of confidence-rating ROCs: A critical test with minimal assumptions. *Psychological Review*, *122*(3), 542–557.

Kellen, D., Klauer, K. C., & Bröder, A. (2013). Recognition memory models and binary-response ROCs: A comparison by minimum description length. *Psychonomic Bulletin & Review*, *20*(4), 693–719.

Kersten, A. W., Earles, J. L., Curtayne, E. S., & Lane, J. C. (2008). Adult age differences in binding actors and actions in memory for events. *Memory & Cognition*, *36*(1), 119–131.

Kilb, A., & Naveh-Benjamin, M. (2007). Paying attention to binding: Further studies assessing the role of reduced attentional resources in the associative deficit of older adults. *Memory and Cognition*, *35*(5), 1162–1174.

Kilb, A., & Naveh-Benjamin, M. (2011). The effects of pure pair repetition on younger and older adults' associative memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(3), 706–719.

Kim, S.-Y., & Giovanello, K. S. (2011). The effects of attention on age-related relational memory deficits: Evidence from a novel attentional manipulation. *Psychology and Aging*, *26*(3), 678–688.

Ko, P. C., Duda, B., Hussey, E., Mason, E., Molitor, R. J., Woodman, G. F., & Ally, B. A. (2014). Understanding age-related reductions in visual working memory capacity: Examining the stages of change detection. *Attention, Perception, &*

Psychophysics, *76*(7), 2015–2030.

Koen, J. D., & Yonelinas, A. P. (2014). The effects of healthy aging, amnestic mild cognitive impairment, and Alzheimer's disease on recollection and familiarity: a meta-analytic review. *Neuropsychology Review*, *24*(3), 332–354.

Kondo, A., & Saiki, J. (2012). Feature-specific encoding flexibility in visual working memory. *PloS ONE*, *7*(12), e50962.

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*(3), 299–312.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573–603.

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Kyllingsbæk, S., & Bundesen, C. (2007). Parallel processing in a multifeature whole-report paradigm. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(1), 64–82.

Lecerf, T., & De Ribaupierre, A. (2005). Recognition in a visuospatial memory task: The effect of presentation. *European Journal of Cognitive Psychology*, *17*(1), 47–75.

Lee, M. D. (2014). *The "new statistics" are built on fundamentally flawed foundations.* Retrieved from `https://webfiles.uci.edu/mdlee/Lee2014_NewStatistics.pdf`

Leonards, U., Ibanez, V., & Giannakopoulos, P. (2002). The role of stimulus type in age-related changes of visual working memory. *Experimental Brain Research*, *146*(2), 172–183.

Li, S. C., & Sikstrom, S. (2002). Integrative neurocomputational perspectives on cognitive aging, neuromodulation, and representation. *Neuroscience and Biobehavioral Reviews*, *26*, 795–808.

Liang, Y., Pertzov, Y., Nicholas, J. M., Henley, S. M., Crutch, S., Woodward, F., ... Husain, M. (2016). Visual short-term memory binding deficit in familial Alzheimer's disease. *Cortex*, *78*, 150–164.

Lin, P.-H., & Luck, S. J. (2012). Proactive interference does not meaningfully distort visual working memory capacity estimates in the canonical change detection task. *Frontiers in psychology*, *3*(42), 1–9.

Link, W. A., & Eaton, M. J. (2012). On thinning of chains in MCMC. *Methods in Ecology and Evolution*, *3*(1), 112–115.

Logie, R. H. (2011). The functional organization and capacity limits of working memory. *Current Directions in Psychological Science*, *20*(4), 240–245.

Logie, R. H., Brockmole, J. R., & Jaswal, S. (2011). Feature binding in visual short-term memory is unaffected by task-irrelevant changes of location, shape, and color. *Memory & Cognition*, *39*(1), 24–36.

Logie, R. H., Brockmole, J. R., & Vandenbroucke, A. R. (2009). Bound feature combinations in visual short-term memory are fragile but influence long-term learning. *Visual Cognition*, *17*(1-2), 160–179.

Logie, R. H., Della Sala, S., Laiacona, M., Chalmers, P., & Wynn, V. (1996). Group aggregates and individual reliability: The case of verbal short-term memory. *Memory & Cognition*, *24*(3), 305–321.

Logie, R. H., Horne, M. J., & Pettit, L. D. (2015). When cognitive performance does not decline across the lifespan. In R. H. Logie & R. G. Morris (Eds.), *Working memory and ageing* (pp. 21–47). Hove, East Sussex: Psychology Press.

Logie, R. H., & Maylor, E. A. (2009). An internet study of prospective memory across adulthood. *Psychology and Aging*, *24*(3), 767–774.

Luce, R. D. (1963a). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1, pp. 103–189). New York, NY: Wiley.

Luce, R. D. (1963b). A threshold theory for simple detection experiments. *Psychological Review*, *70*(1), 61–79.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*, 279–281.

Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: from psychophysics and neurobiology to individual differences. *Trends in Cognitive*

Sciences, *17*(8), 391–400.

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience*, *17*(3), 347–356.

Macmillan, N. A., & Creelman, C. D. (1990). Response bias: Characteristics of detection theory, threshold theory, and "nonparametric" indexes. *Psychological Bulletin*, *107*(3), 401–413.

Macmillan, N. A., & Creelman, C. D. (1996). Triangles in ROC space: History and theory of "nonparametric" measures of sensitivity and response bias. *Psychonomic Bulletin & Review*, *3*(2), 164–170.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Maillet, D., & Schacter, D. L. (2016). From mind wandering to involuntary retrieval: Age-related differences in spontaneous cognitive processes. *Neuropsychologia*, *80*, 142–156.

Makovski, T., & Jiang, Y. V. (2008). Proactive interference from items previously stored in visual working memory. *Memory & Cognition*, *36*(1), 43–52.

Makovski, T., Sussman, R., & Jiang, Y. V. (2008). Orienting attention in visual working memory reduces interference from memory probes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(2), 369–380.

Makovski, T., Watson, L. M., Koutstaal, W., & Jiang, Y. V. (2010). Method matters: systematic effects of testing procedure on visual working memory sensitivity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(6), 1466–1479.

Malmberg, K. J. (2002). On the form of ROCs constructed from confidence ratings. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(2), 380–387.

Mäntylä, T. (1993). Knowing but not remembering: Adult age differences in recollective experience. *Memory & Cognition*, *21*(3), 379–388.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York, NY, USA: Henry Holt

and Co., Inc.

Maylor, E. A., & Logie, R. H. (2010). A large-scale comparison of prospective and retrospective memory development from childhood to middle age. *The Quarterly Journal of Experimental Psychology*, *63*(3), 442–451.

McCabe, D. P., Roediger, H. L., McDaniel, M. A., & Balota, D. A. (2009). Aging reduces veridical remembering but increases false remembering: Neuropsychological test correlates of remember–know judgments. *Neuropsychologia*, *47*(11), 2164–2173.

McIntyre, J. S., & Craik, F. I. M. (1987). Age differences in memory for item and source information. *Canadian Journal of Psychology*, *41*(2), 175–192.

McVay, J. C., Meier, M. E., Touron, D. R., & Kane, M. J. (2013). Aging ebbs the flow of thought: Adult age differences in mind wandering, executive control, and self-evaluation. *Acta psychologica*, *142*(1), 136–147.

Mitchell, K. J., Johnson, M. K., Raye, C. L., & D'Esposito, M. (2000). fMRI evidence of age-related hippocampal dysfunction in feature binding in working memory. *Cognitive Brain Research*, *10*(1), 197–206.

Mitchell, K. J., Johnson, M. K., Raye, C. L., Mather, M., & D'Esposito, M. D. (2000). Aging and reflective processes of working memory: Binding and test load deficits. *Psychology and Aging*, *15*(3), 527–541.

Morey, C. C., & Bieler, M. (2013). Visual short-term memory always requires general attention. *Psychonomic Bulletin and Review*, *20*, 163–170.

Morey, C. C., & Cowan, N. (2004). When visual and verbal memories compete: Evidence of cross-domain limits in working memory. *Psychonomic Bulletin & Review*, *11*(2), 296–301.

Morey, C. C., Morey, R. D., van der Reijden, M., & Holweg, M. (2013). Asymmetric cross-domain interference between two working memory tasks: Implications for models of working memory. *Journal of Memory and Language*, *69*(3), 324–348.

Morey, R. D. (2011). A Bayesian hierarchical model for the measurement of working memory capacity. *Journal of Mathematical Psychology*, *55*(1), 8–24.

Morey, R. D., & Rouder, J. N. (2015). BayesFactor: Computation of Bayes Factors for Common Designs [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=BayesFactor` (R package version 0.9.11-1)

Morey, R. D., Rouder, J. N., & Speckman, P. L. (2008). A statistical model for discriminating between subliminal and near-liminal performance. *Journal of Mathematical Psychology*, *52*(1), 21–36.

Moscovitch, M. (1992). Memory and working-with-memory: A component process model based on modules and central systems. *Journal of Cognitive Neuroscience*, *4*(3), 257–267.

Myerson, J., Emery, L., White, D. A., & Hale, S. (2003). Effects of age, domain, and processing demands on memory span: Evidence for differential decline. *Aging, Neuropsychology, and Cognition*, *10*(1), 20–27.

Myerson, J., Hale, S., Rhee, S. H., & Jenkins, L. (1999). Selective interference with verbal and spatial working memory in young and older adults. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *54*(3), P161–P164.

Naveh-Benjamin, M. (2000). Adult age differences in memory performance: Tests of an associative deficit hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(5), 1170–1187.

Naveh-Benjamin, M., Brav, T. K., & Levy, O. (2007). The associative memory deficit of older adults: the role of strategy utilization. *Psychology and Aging*, *22*(1), 202–208.

Naveh-Benjamin, M., Craik, F. I., Guez, J., & Kreuger, S. (2005). Divided attention in younger and older adults: effects of strategy and relatedness on memory performance and secondary task costs. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 520–537.

Naveh-Benjamin, M., Guez, J., & Shulman, S. (2004). Older adults' associative deficit in episodic memory: Assessing the role of decline in attentional resources. *Psychonomic Bulletin and Review*, *11*(5), 1067–1073.

Naveh-Benjamin, M., Hussain, Z., Guez, J., & Bar-On, M. (2003). Adult age differ-

ences in episodic memory: further support for an associative-deficit hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(5), 826–837.

Naveh-Benjamin, M., Shing, Y. L., Kilb, A., Werkle-Bergner, M., Lindenberger, U., & Li, S.-C. (2009). Adult age differences in memory for name–face associations: The effects of intentional and incidental learning. *Memory*, *17*(2), 220–232.

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, *135*(3), 370–384.

Nelson, H. E. (1982). *National Adult Reading Test (NART): Test manual*. Windsor, UK: NFER-NELSON Publishing.

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature Neuroscience*, *14*(9), 1105–1107.

Nissen, M. J. (1985). Accessing features and objects: Is location special? In M. I. Posner & O. S. M. Marin (Eds.), *Attention and performance xi* (pp. 205–219). Hillsdale, NJ: Erlbaum.

Noack, H., Lovden, M., & Lindenberger, U. (2012). Normal aging increases discriminal dispersion in visuospatial short-term memory. *Psychology and Aging*, *27*(3), 627–637.

Ntzoufras, I. (2009). *Bayesian modeling using WinBUGS* (Vol. 1). John Wiley & Sons.

Nuijten, M. B., Hartgerink, C. H., van Assen, M. A., Epskamp, S., & Wicherts, J. M. (2015). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, 1–22.

Oberauer, K. (2005). Binding and inhibition in working memory: Individual and age differences in short-term recognition. *Journal of Experimental Psychology: General*, *134*(3), 368–387.

Oberauer, K., & Eichenberger, S. (2013). Visual working memory declines when more features must be remembered for each object. *Memory & Cognition*, *41*(8), 1212–1227.

Oberauer, K., Lewandowsky, S., Farrell, S., Jarrold, C., & Greaves, M. (2012). Modeling working memory: An interference model of complex span. *Psychonomic Bulletin & Review*, *19*(5), 779–819.

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Oxford University Press.

Old, S. R., & Naveh-Benjamin, M. (2008a). Differential effects of age on item and associative measures of memory: A meta-analysis. *Psychology and Aging*, *23*(1), 104–118.

Old, S. R., & Naveh-Benjamin, M. (2008b). Memory for people and their actions: further evidence for an age-related associative deficit. *Psychology and Aging*, *23*(2), 467–472.

Olsen, R. K., Moses, S. N., Riggs, L., & Ryan, J. D. (2012). The hippocampus supports multiple cognitive processes through relational binding and comparison. *Frontiers in Human Neuroscience*, *6*.

Olson, I. R., & Jiang, Y. (2002). Is visual short-term memory object based? Rejection of the "strong-object" hypothesis. *Perception and Psychophysics*, *64*(7), 1055–1067.

Olson, I. R., Page, K., Moore, K. S., Chatterjee, A., & Verfaellie, M. (2006). Working memory for conjunctions relies on the medial temporal lobe. *The Journal of Neuroscience*, *26*(17), 4596–4601.

Olson, I. R., Zhang, J. X., Mitchell, K. J., Johnson, M. K., Bloise, S. M., & Higgins, J. A. (2004). Preserved spatial memory over brief intervals in older adults. *Psychology and Aging*, *19*(2), 310–317.

Oosterman, J. M., Morel, S., Meijer, L., Buvens, C., Kessels, R. P., & Postma, A. (2011). Differential age effects on spatial and visual working memory. *The International Journal of Aging and Human Development*, *73*(3), 195–208.

Parra, M. A. (2014). Overcoming barriers in cognitive assessment of Alzheimer's disease. *Dementia & Neuropsychologia*, *8*(2), 95–98.

Parra, M. A., Abrahams, S., Fabi, K., Logie, R. H., Luzzi, S., & Della Sala, S. (2009). Short-term memory binding deficits in Alzheimer's disease. *Brain*,

132, 1057–1066.

Parra, M. A., Abrahams, S., Logie, R. H., & Della Sala, S. (2009). Age and binding within-dimension features in visual short-term memory. *Neuroscience Letters*, *449*, 1–5.

Parra, M. A., Abrahams, S., Logie, R. H., & Della Sala, S. (2010). Visual short-term memory binding in Alzheimer's disease and depression. *Journal of Neurology*, *257*(7), 1160–1169.

Parra, M. A., Abrahams, S., Logie, R. H., Mendez, L. G., Lopera, F., & Della Sala, S. (2010). Visual short-term memory binding deficits in familial Alzheimer's disease. *Brain*, *133*, 2702–2713.

Parra, M. A., Cubelli, R., & Della Sala, S. (2011). Lack of color integration in visual short-term memory binding. *Memory and Cognition*, *39*, 1187–1197.

Parra, M. A., Della Sala, S., Abrahams, S., Logie, R. H., Mendez, L. G., & Lopera, F. (2011). Specific deficit of colour-colour short-term memory binding in sporadic and familial Alzheimer's disease. *Neuropsychologia*, *49*, 1943–1952.

Parra, M. A., Della Sala, S., Logie, R. H., & Morcom, A. M. (2014). Neural correlates of shape–color binding in visual working memory. *Neuropsychologia*, *52*, 27–36.

Parra, M. A., Fabi, K., Luzzi, S., Cubelli, R., Hernandez Valdez, M., & Della Sala, S. (2015). Relational and conjunctive binding functions dissociate in short-term memory. *Neurocase*, *21*(1), 56–66.

Parra, M. A., Ibáñez, A., Rhodes, S., Baglivo, F., Kargieman, L., García, M. C., ... Della Sala, S. (under review). *Behavioural and electrophysiological correlates of visual working memory capacity during change detection tasks.*

Parra, M. A., Saarimäki, H., Bastin, M. E., Londoño, A. C., Pettit, L., Lopera, F., ... Abrahams, S. (2015). Memory binding and white matter integrity in familial alzheimer's disease. *Brain*, online ahead of print.

Pashler, H. (1988). Familiarity and visual change detection. *Perception and Psychophysics*, *44*(4), 369–378.

Pastore, R. E., Crawley, E. J., Berens, M. S., & Skelly, M. A. (2003). "Nonpara-

metric" A' and other modern misconceptions about signal detection theory. *Psychonomic Bulletin & Review*, *10*(3), 556–569.

Pazzaglia, A. M., Dube, C., & Rotello, C. M. (2013). A critical comparison of discrete-state and continuous models of recognition memory: Implications for recognition and beyond. *Psychological Bulletin*, *139*(6), 1173–1203.

Peich, M.-C., Husain, M., & Bays, P. M. (2013). Age-related decline of precision and binding in visual working memory. *Psychology and Aging*, *28*(3), 729–743.

Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1), 8–13.

Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics*, *2*, 10. doi: 10.3389/neuro.11.010.2008

Perfect, T. J., & Maylor, E. A. (2000). Rejecting the dull hypothesis: The relation between method and theory in cognitive aging research. In T. J. Perfect & E. A. Maylor (Eds.), *Models of cognitive aging* (pp. 1–18). Oxford, UK: Oxford University Press.

Pertzov, Y., Dong, M. Y., Peich, M.-C., & Husain, M. (2012). Forgetting what was where: The fragility of object-location binding. *PLoS ONE*, *7*(10), e48214. doi: 10.1371/journal.pone.0048214

Pertzov, Y., Heider, M., Liang, Y., & Husain, M. (2015). Effects of healthy ageing on precision and binding of object location in visual short term memory. *Psychology and Aging*, *30*(1), 26–35.

Pertzov, Y., Miller, T. D., Gorgoraptis, N., Caine, D., Schott, J. M., Butler, C., & Husain, M. (2013). Binding deficits in memory following medial temporal lobe damage in patients with voltage-gated potassium channel complex antibody-associated limbic encephalitis. *Brain*, *136*(8), 2474–2485.

Peterson, D. J., & Naveh-Benjamin, M. (2016). The role of aging in intra-item and item-context binding processes in visual working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, on-line ahead of print.

Phillips, W. A. (1974). On the distinction between sensory storage and short-term

visual memory. *Perception and Psychophysics*, *16*(2), 283–290.

Piekema, C., Kessels, R. P., Mars, R. B., Petersson, K. M., & Fernández, G. (2006). The right hippocampus participates in short-term memory maintenance of object–location associations. *NeuroImage*, *33*(1), 374–382.

Piekema, C., Kessels, R. P., Rijpkema, M., & Fernández, G. (2009). The hippocampus supports encoding of between-domain associations within working memory. *Learning and Memory*, *16*(4), 231–234.

Piekema, C., Rijpkema, M., Fernández, G., & Kessels, R. P. (2010). Dissociating the neural correlates of intra-item and inter-item working-memory binding. *PloS ONE*, *5*(4), e10214.

Pinto, Y., Sligte, I. G., Shapiro, K. L., & Lamme, V. A. (2013). Fragile visual short-term memory is an object-based and location-specific store. *Psychonomic Bulletin & Review*, *20*(4), 732–739.

Plummer, M. (2002). Discussion of the paper by Spiegelhalter et al. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 620.

Plummer, M. (2015). rjags: Bayesian Graphical Models using MCMC [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=rjags` (R package version 3-15)

Plummer, M., et al. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing* (Vol. 124, p. 125).

Pollack, I., & Hsieh, R. (1969). Sampling variability of the area under the ROC-curve and of d'_e. *Psychological Bulletin*, *71*(3), 161.

Pollack, I., & Norman, D. A. (1964). A non-parametric analysis of recognition experiments. *Psychonomic Science*, *1*(1-12), 125–126.

Postma, A., & De Haan, E. H. F. (1996). What was where? Memory for object locations. *The Quarterly Journal of Experimental Psychology*, *49A*(1), 178–199.

Postma, A., Kessels, R. P., & van Asselen, M. (2008). How the brain remembers and forgets where things are: The neurocognition of object–location memory.

Neuroscience & Biobehavioral Reviews, *32*(8), 1339–1345.

Prinzmetal, W., & Keysar, B. (1989). Functional theory of illusory conjunctions and neon colors. *Journal of Experimental Psychology: General*, *118*(2), 165–190.

Province, J. M., & Rouder, J. N. (2012). Evidence for discrete-state processing in recognition memory. *Proceedings of the National Academy of Sciences*, *109*(36), 14357–14362.

Quinlan, P. T. (2003). Visual feature integration theory: past, present, and future. *Psychological Bulletin*, *129*(5), 643–673.

R Core Team. (2015). R: A Language and Environment for Statistical Computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Ranganath, C. (2010). Binding items and contexts the cognitive neuroscience of episodic memory. *Current Directions in Psychological Science*, *19*(3), 131–137.

Ratcliff, R., Sheu, C.-F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review*, *99*(3), 518–535.

Raz, N., & Rodrigue, K. M. (2006). Differential aging of the brain: patterns, cognitive correlates and modifiers. *Neuroscience & Biobehavioral Reviews*, *30*(6), 730–748.

Read, C. A., Rogers, J. M., & Wilson, P. H. (2016). Working memory binding of visual object features in older adults. *Aging, Neuropsychology, and Cognition*, *23*(3), 263–281.

Redel, P., Bublak, P., Sorg, C., Kurz, A., Förstl, H., Müller, H., ... Finke, K. (2012). Deficits of spatial and task-related attentional selection in mild cognitive impairment and Alzheimer's disease. *Neurobiology of Aging*, *33*(1), e27–e42.

Rhodes, M. G., Castel, A. D., & Jacoby, L. L. (2008). Associative recognition of face pairs by younger and older adults: the role of familiarity-based processing. *Psychology and Aging*, *23*(2), 239–249.

Rhodes, S., & Parra, M. A. (2016). Executive functioning. In N. A. Pachana (Ed.), *Encyclopedia of geropsychology.* Singapore: Springer.

Rhodes, S., Parra, M. A., & Logie, R. H. (2016). Ageing and feature binding in visual working memory: The role of presentation time. *The Quarterly Journal of Experimental Psychology*, *69*(4), 654–668. doi: 10.1080/17470218 .2015.1038571

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, *95*(3), 318–339.

Rotello, C. M., Heit, E., & Dubé, C. (2015). When more data steer us wrong: Replications with the wrong dependent measure perpetuate erroneous conclusions. *Psychonomic Bulletin & Review*, *22*, 944–954.

Rotello, C. M., Masson, M. E., & Verde, M. F. (2008). Type I error rates and power analyses for single-point sensitivity measures. *Perception & Psychophysics*, *70*(2), 389–401.

Rouder, J. N., Morey, R. D., Cowan, N., Zwilling, C. E., Morey, C. C., & Pratte, M. S. (2008). An assessment of fixed-capacity models of visual working memory. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(16), 5975–5979.

Rouder, J. N., Morey, R. D., Morey, C. C., & Cowan, N. (2011). How to measure working memory capacity in the change detection paradigm. *Psychonomic Bulletin and Review*, *18*, 324–330.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Pratte, M. S. (2007). Detecting chance: A solution to the null sensitivity problem in subliminal priming. *Psychonomic Bulletin & Review*, *14*(4), 597–605.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*, 356–374.

Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., Wagenmakers, E.-J., & Rouder, J. (submitted). The $p < .05$ rule and the hidden costs of the free lunch in inference. *Manuscript under review*.

Rouder, J. N., Morey, R. D., Verhagen, J., Swagman, A. R., & Wagenmakers, E.-J. (in press). Bayesian analysis of factorial designs.

Rouder, J. N., Province, J. M., Swagman, A. R., & Thiele, J. E. (submitted). From ROC curves to psychological theory. *Manuscript submitted for publication*.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review*, *16*(2), 225–237.

Sala, J. B., & Courtney, S. M. (2007). Binding of what and where during working memory maintenance. *Cortex*, *43*(1), 5–21.

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review*, *103*(3), 403–428.

Salthouse, T. A. (2000). Methodological assumptions in cognitive aging research. In F. I. M. Craik & T. A. Salthouse (Eds.), *The handbook of cognitive aging* (2nd ed., pp. 467–498). Mahwah, NJ: Erlbaum.

Sander, M. C., Lindenberger, U., & Werkle-Bergner, M. (2012). Lifespan age differences in working memory: A two-component framework. *Neuroscience and Biobehavioral Reviews*, *36*, 2007–2033.

Sander, M. C., Werkle-Bergner, M., & Lindenberger, U. (2011a). Binding and strategic selection in working memory: A lifespan dissociation. *Psychology and Aging*, *26*(3), 612–624.

Sander, M. C., Werkle-Bergner, M., & Lindenberger, U. (2011b). Contralateral delay activity reveals life-span age differences in top-down modulation of working memory contents. *Cerebral Cortex*, *21*(12), 2809–2819.

Sander, M. C., Werkle-Bergner, M., & Lindenberger, U. (2012). Amplitude modulations and inter-trial phase stability of alpha-oscillations differentially reflect working memory constraints across the lifespan. *Neuroimage*, *59*(1), 646–654.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-prime: User's guide*. Psychology Software Incorporated.

Schooler, L. J., & Shiffrin, R. M. (2005). Efficiently measuring recognition performance with sparse data. *Behavior Research Methods*, *37*(1), 3–10.

Sellke, T., Bayarri, M., & Berger, J. O. (2001). Calibration of ρ values for testing precise null hypotheses. *The American Statistician*, *55*(1), 62–71.

Shing, Y. L., Werkle-Bergner, M., Brehmer, Y., Muller, V., Li, S.-C., & Lindenberger, U. (2010). Episodic memory across the lifespan: The contribution of associative and strategic components. *Neuroscience and Biobehavioral Reviews*, *34*, 1080–1091.

Shing, Y. L., Werkle-Bergner, M., Li, S.-C., & Lindenberger, U. (2008). Associative and strategic components of episodic memory: a life-span dissociation. *Journal of Experimental Psychology: General*, *137*(3), 495–513.

Shipstead, Z., & Engle, R. W. (2013). Interference within the focus of attention: Working memory tasks reflect more than temporary maintenance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(1), 277–289.

Sims, C. R., Jacobs, R. A., & Knill, D. C. (2012). An ideal observer analysis of visual working memory. *Psychological Review*, *119*(4), 807–830.

Sligte, I. G., Scholte, H. S., & Lamme, V. A. (2008). Are there multiple visual short-term memory stores? *PLOS one*, *3*(2), e1699.

Smith, W. D. (1995). Clarification of sensitivity measure A'. *Journal of Mathematical Psychology*, *39*(1), 82–89.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*(1), 34–50.

Song, J.-H., & Jiang, Y. (2006). Visual working memory for simple and complex features: An fMRI study. *NeuroImage*, *30*(3), 963–972.

Spencer, W. D., & Raz, N. (1995). Differential effects of aging on memory for content and context: a meta-analysis. *Psychology and Aging*, *10*(4), 527–539.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied*, *74*(11), 1–29.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, and Computers*, *31*(1), 137–

149.

Stefurak, D. L., & Boynton, R. M. (1986). Independence of memory for categorically different colors and shapes. *Perception & Psychophysics*, *39*(3), 164–174.

Störmer, V. S., Passow, S., Biesenack, J., & Li, S.-C. (2012). Dopaminergic and cholinergic modulations of visual-spatial attention and working memory: Insights from molecular genetic research and implications for adult cognitive development. *Developmental Psychology*, *48*(3), 875–889.

Suchow, J. W., Fougnie, D., Brady, T. F., & Alvarez, G. A. (2014). Terms of the debate on the format and structure of visual memory. *Attention, Perception, & Psychophysics*, *76*(7), 2071–2079.

Swets, J. A. (1986a). Form of empirical ROCs in discrimination and diagnostic tasks: implications for theory and measurement of performance. *Psychological Bulletin*, *99*(2), 181–198.

Swets, J. A. (1986b). Indices of discrimination or diagnostic accuracy: their ROCs and implied models. *Psychological Bulletin*, *99*(1), 100–117.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240*(4857), 1285–1293.

Tanner Jr, W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, *61*(6), 401–409.

Treisman, A. (1977). Focused attention in the perception and retrieval of multidimensional stimuli. *Perception & Psychophysics*, *22*(1), 1–11.

Treisman, A. (1996). The binding problem. *Current Opinion in Neurobiology*, *6*(2), 171–178.

Treisman, A. (2006). Object tokens, binding, and visual memory. In H. D. Zimmer, A. Mecklinger, & U. Lindenberger (Eds.), *Handbook of binding and memory: perspectives from cognitive neuroscience* (pp. 315–338). New York, NY: Oxford University Press.

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.

Treisman, A., & Schmidt, H. (1982). Illusory conjunctions in the perception of

objects. *Cognitive Psychology*, *14*(1), 107–141.

Treisman, A., Sykes, M., & Gelade, G. (1977). Selective attention and stimulus integration. In *Attention and Performance VI* (pp. 333–361). Hillsdale, NJ: Erlbaum.

Treisman, A., & Zhang, W. (2006). Location and binding in visual working memory. *Memory & Cognition*, *34*(8), 1704–1719.

Troyer, A. K., & Craik, F. I. M. (2000). The effect of divided attention on memory for items and their context. *Canadian Journal of Experimental Psychology*, *54*(3), 161–171.

Troyer, A. K., Winocur, G., Craik, F. I. M., & Moscovitch, M. (1999). Source memory and divided attention: Reciprocal costs to primary and secondary tasks. *Neuropsychology*, *13*(4), 467–474.

Tse, C.-S., Balota, D. A., Moynan, S. C., Duchek, J. M., & Jacoby, L. L. (2010). The utility of placing recollection in opposition to familiarity in early discrimination of healthy aging and very mild dementia of the Alzheimer's type. *Neuropsychology*, *24*(1), 49–67.

Tubi, N., & Calev, A. (1989). Verbal and visuospatial recall by younger and older subjects: use of matched tasks. *Psychology and Aging*, *4*(4), 493–495.

Tulving, E. (1972). Episodic and semantic memory. In D. W. Tulving E (Ed.), *Organization of memory* (pp. 381–403). New York, NY: Academic Press.

Ueno, T., Allen, R. J., Baddeley, A. D., Hitch, G. J., & Saito, S. (2011). Disruption of visual feature binding in working memory. *Memory and Cognition*, *39*, 12–23.

Ueno, T., Mate, J., Allen, R. J., Hitch, G. J., & Baddeley, A. D. (2011). What goes through the gate? Exploring interference with visual feature binding. *Neuropsychologia*, *49*, 1597–1604.

Unsworth, N., & Robison, M. K. (2016). The influence of lapses of attention on working memory capacity. *Memory & Cognition*, *44*, 188–196.

van den Berg, R., Awh, E., & Ma, W. J. (2014). Factorial comparison of working memory models. *Psychological Review*, *121*(1), 124–149.

van den Berg, R., Shin, H., Chou, W.-C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences*, *109*(22), 8780–8785.

van Geldorp, B., Parra, M. A., & Kessels, R. P. (2015). Cognitive and neuropsychological underpinnings of relational and conjunctive working memory binding across age. *Memory*, *23*(8), 1112–1122.

Van Snellenberg, J. X., Conway, A. R., Spicer, J., Read, C., & Smith, E. E. (2014). Capacity estimates in working memory: Reliability and interrelationships among tasks. *Cognitive, Affective, & Behavioral Neuroscience*, *14*(1), 106–116.

Verhaeghen, P. (2011). Aging and executive control: reports of a demise greatly exaggerated. *Current Directions in Psychological Science*, *20*(3), 174–180.

Verhaeghen, P., Marcoen, A., & Goossens, L. (1993). Facts and fiction about memory aging: A quantitative integration of research findings. *Journal of Gerontology*, *48*(4), P157–P171.

Verhaeghen, P., & Salthouse, T. A. (1997). Meta-analyses of age–cognition relations in adulthood: Estimates of linear and nonlinear age effects and structural models. *Psychological bulletin*, *122*(3), 231–249.

Vogel, E. K., & Awh, E. (2008). How to exploit diversity for scientific gain using individual differences to constrain cognitive theory. *Current Directions in Psychological Science*, *17*(2), 171–176.

Vogel, E. K., & Machizawa, M. G. (2004). Neural activity predicts individual differences in visual working memory capacity. *Nature*, *428*(6984), 748–751.

Vogel, E. K., McCollough, A. W., & Machizawa, M. G. (2005). Neural measures reveal individual differences in controlling access to working memory. *Nature*, *438*(7067), 500–503.

Vul, E., Hanus, D., & Kanwisher, N. (2009). Attention as inference: selection is probabilistic; responses are all-or-none samples. *Journal of Experimental Psychology: General*, *138*(4), 546–560.

Vul, E., & Rich, A. N. (2010). Independent sampling of features enables conscious

perception of bound objects. *Psychological Science*, *21*(8), 1168–1175.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804.

Walker, P., & Cuthbert, L. (1998). Remembering visual feature conjunctions: visual memory for shape-colour associations is object-based. *Visual Cognition*, *5*(4), 409–455.

West, R. (1999). Visual distraction, working memory, and aging. *Memory & Cognition*, *27*(6), 1064–1072.

West, R. L. (1996). An application of prefrontal cortex function theory to cognitive aging. *Psychological Bulletin*, *120*(2), 272–292.

Wheeler, M. E., & Treisman, A. M. (2002). Binding in short-term visual memory. *Journal of Experimental Psychology: General*, *131*(1), 48–64.

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision*, *4*(12), 1120–1135.

Wilton, R. N. (1989). The structure of memory: Evidence concerning the recall of surface and background colour of shapes. *The Quarterly Journal of Experimental Psychology*, *41*(3), 579–598.

Wolfe, J. M., & Cave, K. R. (1999). The psychophysical evidence for a binding problem in human vision. *Neuron*, *24*(1), 11–17.

Woodman, G. F., & Vogel, E. K. (2008). Selective storage and maintenance of an object's features in visual working memory. *Psychonomic Bulletin & Review*, *15*(1), 223–229.

Woodman, G. F., Vogel, E. K., & Luck, S. J. (2012). Flexibility in visual working memory: Accurate change detection in the face of irrelevant variations in position. *Visual Cognition*, *20*(1), 1–28.

Xu, Y. (2002). Encoding color and shape from different parts of an object in visual short-term memory. *Perception & Psychophysics*, *64*(8), 1260–1280.

Xu, Y. (2007). The role of the superior intraparietal sulcus in supporting visual short-term memory for multifeature objects. *The Journal of Neuroscience*, *27*(43), 11676–11686.

Xu, Y., & Chun, M. M. (2006). Dissociable neural mechanisms supporting visual short-term memory for objects. *Nature*, *440*(7080), 91–95.

Yang, C.-T., Tseng, P., & Wu, Y.-J. (2015). Effect of decision load on whole-display superiority in change detection. *Attention, Perception, & Psychophysics*, *77*(3), 749–758.

Yeh, Y.-Y., Yang, C.-T., & Chiu, Y.-C. (2005). Binding or prioritization: The role of selective attention in visual short-term memory. *Visual Cognition*, *12*, 759–799.

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: a review. *Psychological Bulletin*, *133*(5), 800–832.

Zavagnin, M., Borella, E., & De Beni, R. (2014). When the mind wanders: Age-related differences between young and older adults. *Acta psychologica*, *145*, 54–64.

Zeki, S. (1976). The functional organization of projections from striate to prestriate visual cortex in the rhesus monkey. In *Cold spring harbor symposia on quantitative biology* (Vol. 40, pp. 591–600).

Zhang, H., Xuan, Y., Fu, X., & Pylyshyn, Z. W. (2010). Do objects in working memory compete with objects in perception? *Visual Cognition*, *18*(4), 617–640.

Zhang, J., & Mueller, S. T. (2005). A note on ROC analysis and non-parametric estimate of sensitivity. *Psychometrika*, *70*(1), 203–212.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233–235.

Zimmer, H. D., & Ecker, U. K. (2010). Remembering perceptual features unequally bound in object and episodic tokens: Neural mechanisms and their electrophysiological correlates. *Neuroscience & Biobehavioral Reviews*, *34*(7), 1066–1079.

Zimmer, H. D., Mecklinger, A., & Lindenberger, U. (2006). Levels of binding: types, mechanisms, and functions of binding in remembering. In H. D. Zimmer, A. Mecklinger, & U. Lindenberger (Eds.), *Handbook of binding and memory:*

perspectives from cognitive neuroscience (pp. 3–22). New York, NY: Oxford University Press.

Zokaei, N., Heider, M., & Husain, M. (2014). Attention is required for maintenance of feature binding in visual working memory. *The Quarterly Journal of Experimental Psychology*, *67*(6), 1191–1213.

Appendix A

Hierarchical Logit Model

A.1 Detailed Description of the Model

In this analysis the log odds of a correct response on a given trial was modelled as a linear combination of a grand mean parameter and deflections from the grand mean that represent main and interaction effects of our experimental factors. These deflections were constrained to sum-to-zero via the use of effects coded variables (Ntzoufras, 2009). In effects coding, as with many other coding schemes, we are limited to $I-1$ indicator variables, where I is the number of levels in a given factor. One level is set to -1 for all indicator variables and acts as the reference level; the $I-1$ variables then reflect the deflection from the mean attributable to each remaining level with that level coded 1 and the rest (except for the reference level) coded 0. The resulting coefficient associated with an indicator variable reflects the deflection from the mean associated with the positively coded factor level. These coefficients are constrained to sum-to-zero and the corresponding coefficient for the reference level is the negative sum of the $I-1$ coefficients associated with a given factor. Interaction variables are analysed similarly and reflect the product of these effects coded indicator variables (see, Ntzoufras, 2009). The coding schemes used to create the design matrix ($\mathbf{X}$) for each analysis are reported with the Table of posterior quantities.

Finally, as we had repeated measures from the same individuals across conditions,

we modelled an additional effect of participant reflecting the fact that individuals will vary in their overall level of performance. Participant effects were assumed to be drawn from a normal distribution with a mean of zero and a standard deviation estimated from the data (as is typical in hierarchical modelling. For example see, Gelman & Hill, 2007). The full model can be summaraised as follows:

$$
\begin{aligned}
y_i &\sim \text{Bernoulli}(\pi_i), \quad \text{for } i \text{ in } 1, \ldots, t \\
\text{logit}(\pi_i) &= \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \mathbf{X}_i\boldsymbol{\beta} + s_{j[i]} \\
s_j &\sim \text{Normal}(0,\, \sigma_s), \quad \text{for } j \text{ in } 1, \ldots, n
\end{aligned}
$$

where t is the number observations (or trials) and n is the number of participants[1]. The first line gives the likelihood function; each trial is assumed to be a Bernoulli random variable with the underlying probability of success, π_i, determined by the second line. This second line models the log odds of the underlying success probability parameter as a linear combination of three components; (1) a grand mean parameter, β_0, (2) deflections from the grand mean represented by the parameter vector, $\boldsymbol{\beta}$, which is multiplied by the row in the matrix $\mathbf{X}$ containing the effects coded indicator variables for the corresponding trial, and (3) an additional participant level effect. The final line reflects the assumption that participant effects are normally distributed with a mean of 0 and standard deviation, σ_s.

As our estimation is Bayesian, prior distributions must be placed on model parameters. For fixed effects we follow the suggestions of Gelman et al. (2008) and use a mildly informative Cauchy prior:

$$
\beta_l \sim \text{Cauchy}(0,\, 2.5), \quad \text{for } l \text{ in } 1, \ldots, p
$$

where p is the number of effects-coded variables, which in this case was 11. This mildly informative distribution reflects the belief that effects on the log odds scale will usually fall within a restricted range ($\pm$ 2.5) but, due to the Cauchy's heavy tails, does not rule out the possibility of larger effects. A grand mean of 0 in log

[1]For the analysis of Experiments 1 and 2 in Chapter 2 n is 24 and consequently the number of observations, t, in each data set is 10368 as there are 36 trials in each memory condition (3) at each set size (2) for each task (2), resulting in 432 experimental trials for each of the 24 participants.

odds space implies that average performance is at chance, therefore, to reflect our prior expectation that overall performance is likely to be above chance, our prior on β_0 was also a Cauchy distribution centered at 1 (corresponding to approximately 0.73 in probability space) with scale of 2.5. Finally, as is common in Bayesian hierarchical modelling, we place a Gamma prior on the standard deviation of the random participant effect:

$$\tau_s \sim \text{Gamma}(1.01005,\ 0.1005012).$$

This distribution has a mode of 0.1 and standard deviation of 10 (see, Kruschke, 2015), thus is sufficiently broad on the log odds scale.

We took 50000 samples from the posterior distribution across 4 independent MCMC chains using JAGS (Just Another Gibbs Sampler, Plummer et al., 2003) after a burn-in period of 5000 samples. The JAGS model code is given below. A multivariate BGR statistic of 1 was taken to indicate that the chains had converged on a stable distribution (Brooks & Gelman, 1998). It is common to thin MCMC chains (i.e. only retain every k^{th} sample) to reduce auto-correlation, however following the suggestions of Link and Eaton (2012) we do not do this and instead retain the whole large sample, which is more representative of the true posterior distribution than a smaller, thinned chain. Further, we ensured that the *effective sample size* (ESS, Kass et al., 1998), the number of independent samples accounting for autocorrelation, was at least 10000 for the deflection parameters (as per the recommendations of Kruschke, 2015). The deflection parameters (contained in $\boldsymbol{\beta}$) are of primary interest and indicate the size of effects/ interactions in the data, thus we use the resulting posterior samples of these coefficients to construct specific contrasts that test hypotheses about patterns of performance.

A.2 JAGS Model Code

```
model {
  for (i in 1:n){
    y[i] ~ dbern(y.hat[i])
    y.hat[i] <- max(0, min(1,P[i]))
    logit(P[i]) <- B0 + inprod(B, X[i,]) + s[id[i]]
  }
  # grand mean
  B0 ~ dt(1, 1/2.5^2, 1) # places most mass over ~.7
  # deflections from grand mean (fixed effects)
  for (b in 1:nEff){
    B[b] ~ dt(0, 1/2.5^2, 1) # cauchy(0, 2.5) prior (Gelman et al., 2008)
  }
  # participant random effect
  for (ppt in 1:S){
    s[ppt] ~ dnorm(0, sTau)
  }
  sTau <- 1 / pow( sSD , 2 )
  sSD ~ dgamma(1.01005, 0.1005012) # mode = .1, SD = 10 (v. vauge)
}
```

A.3 Interpreting Effects on Log Odds Scale

Throughout the present work we use a recognition paradigm in which the data generated is inherently binary (correct/ incorrect). The rationale in behind the use of a logit model is clearly outlined in Chapter 1, however, the interpretation of coefficients and contrasts (i.e. *effects*) on a log odds scale deserves elaboration. Here we attempt to place effects in log odds space in approximate relation to effects in probability space.

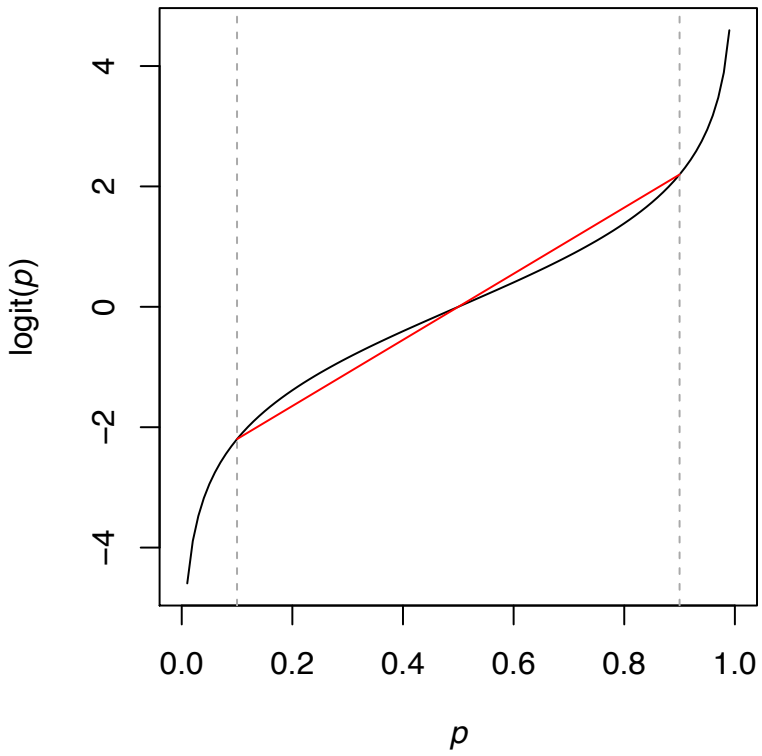

Figure A.1: The relation between probability and log odds with an approximate linear relationship between 0.1 and 0.9 in probability space.

Figure A.1 shows the relationship between probability of success and the log odds of success. This is clearly a non-linear relationship, however restricting our consideration to probabilities between 0.1 and 0.9 we can approximate a linear relationship between p and logit(p) (red line in Figure A.1). In our experiments average performance is always above chance (0.5) and rarely at ceiling (> 0.9) so focusing on this range to define effect sizes is reasonable.

The slope of the red line in Figure A.1 implies that per unit change in p there is approximately a 5.493 change in log odds. Taking the inverse of this slope we find that this approximate linear relationship implies a change of 0.182 in p per unit change in log odds. Using this relationship we can approximately map effects of various sizes in probability space onto the log odds space of our model coefficients. To do this we chose 5 effects in probability space ranging from very small (0.01) to very

Table A.1: Mapping probability effects to log odds via approximate linear relationship

Verbal Category	p	logit(p)
very small	0.01	0.06
small	0.05	0.28
medium	0.10	0.55
large	0.20	1.10
very large	0.30	1.65

large (0.3). Of course selection of these 5 values—and the verbal labels attributed to them—are arbitrary and provide a rough guide to interpretation. Table A.1 presents these 5 effects and their corresponding approximate log odds effect size.

As Table A.1 shows, small effects on the probability scale—that is effects around 0.05 and smaller—are associated with effects on the log odds scale of around 0.3 or smaller. Larger effects that may have more practical relevance—say, effects in the order of 0.1 or larger on the probability scale—are implied by a log odds effect of approximately 0.5 or greater. As acknowledged above this is an approximation for probabilities in the range of 0.1 to 0.9. As Figure A.1 shows this approximation badly breaks down beyond these limits. Further, what is considered a 'small' or 'large' effect in terms of probability will vary from researcher-to-researcher, but Table A.1 provides a rough heuristic.

Appendix B

Additional Detail on ROC Curves

B.1 Constructing a ROC curve

To create our empirical ROCs (see Figure 6.2) the 6 response categories (3 for *same* and 3 for *different* responses) were rearranged to go from 1, representing a high confidence *different* response, to 6, representing a high confidence *same* response. The hit and false alarm pairs that define the ROC are then calculated from the frequency with which each of the 6 ratings is given conditional on whether the probe was in fact same or different (Yonelinas & Parks, 2007). The first point of the ROC curve is given by rating 6 (high confidence *same*) as is estimated via, $(f_6, h_6) = (\Pr[6 \mid \text{different}], \Pr[6 \mid \text{same}])$. The second point builds cumulatively on the first, $(f_5, h_5) = (\Pr[5 \mid \text{different}] + \Pr[5 \mid \text{different}], \Pr[5 \mid \text{same}] + \Pr[5 \mid \text{same}])$ and so on until the full curve is drawn.

B.2 Estimating area under the ROC curve

In the present work we calculated two estimates of area; one parametric, which was the focus of our analysis, and one that makes explicit assumptions regarding the recognition process.

Non-Parametric Area – A_g

Using a confidence rating procedure it is possible to obtain truly non-parametric estimates of sensitivity and A_g is one such measure. As discussed above the 6 points of the ROC curve create trapezoids when connected to the x-axis and the estimate of area is the sum of these areas. Macmillan and Creelman (2005) provide the formula for this on page 64 of their book,

$$A_g = \frac{1}{2}\sum_{i=1}^{6}(F_{i+1} - F_i)(H_{i+1} + H_i),$$

where F and H are vectors containing the 7 false-alarm and hit points along our ROC, respectively, and both start at 0 and end at 1 (see Figure 6.2).

SDT Area – A_z

The measure A_z is derived from SDT and assumes that the underlying evidence distributions are normal but do not necessarily share variance. The normality assumption means that hit and false alarm rates can be converted to z-scores, where according to SDT they should exhibit a linear relationship. Estimating the intercept and slope of this relationship in z-space provides an estimate of the area under the curve, hence the name A_z (Swets, 1988).

To do this we followed the instructions of Stanislaw and Todorov (1999) and conduced two OLS regressions; one regressing hit rate onto false alarm rate and another regressing false-alarm rate onto hit rate, in order to obtain two slopes, $s_1 = \frac{\Delta z(h)}{\Delta z(f)}$ and $s_2 = \frac{\Delta z(f)}{\Delta z(h)}$. The compromise between these two slopes gives us an unbiased estimate of the slope, $S = \frac{1}{2}(s_1 + \frac{1}{s_2})$. The intercept is found as follows, $I = m - (Sn)$, where m is the mean of the 5 hit z-scores and n is the mean of the false-alarm z-scores. The estimate of area is then given by,

$$A_z = \Phi\left(\frac{I}{\sqrt{1+S^2}}\right).$$

Appendix C

JAGS Code for Exploratory Modelling

```
model {
  for (i in 1:n){ # each trial
    y[i] ~ dbern(y.hat[i]) # is a Bernoulli trial
    y.hat[i] <- max(0, min(1, P[i]))

    # determining probability of success... (SS = set size)
    d[i] <- min(1, k[i]/SS[i])
    c[i] <- ifelse(k[i] < (SS[i] - 1),
            1 - (((SS[i] - k[i])*(SS[i] - k[i] - 1))/(SS[i]*(SS[i] - 1))),
            1)
    # informed guessing
    g_f[i] <- u[i]/(u[i] + (1 - d[i])*(1 - u[i]))
    g_b[i] <- ((1 - c[i])*u[i])/((1 - c[i])*u[i] + (1 - d[i])*(1 - u[i]))

    # mcB = 1 if memory condition = binding
    # ttC = 1 if probe type = change (different)
    P[i] <- mcB[i]*(ttC[i]*SP.bc[i] + abs(1 - ttC[i])*SP.bnc[i]) +
            abs(1 - mcB[i])*(ttC[i]*SP.fc[i] + abs(1 - ttC[i])*SP.fnc[i])
```

```
  SP.fc[i] <- a[i]*g_f[i] + (1 - a[i])*u[i]
  SP.fnc[i] <- 1 - (a[i]*(1 - d[i])*g_f[i] + (1 - a[i])*u[i])
  SP.bc[i] <- a[i]*(c[i] + (1 - c[i])*g_b[i]) + (1 - a[i])*u[i]
  SP.bnc[i] <- 1 - (a[i]*(1 - d[i])*g_b[i] + (1 - a[i])*u[i])

  # model transformations of k and g
  k[i] <- max(kappa[i], 0) # Mass-at-chance transformation
  kappa[i] <- K.mu + inprod(K, X_k[i,]) + K_s[id[i]]
  logit(u[i]) <- U.mu + inprod(U, X_u[i,]) + U_s[id[i]]
  logit(a[i]) <- A.mu + inprod(A, X_a[i,]) + A_s[id[i]]
  # each parameter determined by:
  # 1) grand mean
  # 2) STZ effects (deflections from mean - parameters of interest)
  # 3) random ppt effect
}
# Priors
# needed for grand means, STZ effects, and SD of participant effects
### Grand Means
K.mu ~ dnorm(2.5, 1/10^2)
U.mu ~ dnorm(0, 1/10^2)
A.mu ~ dnorm(3, 1/10^2)

### STZ effects
# capacity
for (k.eff in 1:nKeff){
  K[k.eff] ~ dnorm(0, 1/10^2)
}
# guessing
for (u.eff in 1:nUeff){
```

```
    U[u.eff] ~ dnorm(0, 1/10^2)
  }
  # attention
  for (a.eff in 1:nAeff){
    A[a.eff] ~ dnorm(0, 1/10^2)
  }

  ### SD ppt effects
  for (s in 1:S){
    K_s[s] ~ dnorm(0, K_Tau)
    U_s[s] ~ dnorm(0, U_Tau)
    A_s[s] ~ dnorm(0, A_Tau)
  }
  K_Tau <- 1 / pow(K_SD, 2 )
  K_SD ~ dgamma(1.01005, 0.1005012) # mode = .1, SD = 10 (v. vauge)
  U_Tau <- 1 / pow(U_SD, 2 )
  U_SD ~ dgamma(1.01005, 0.1005012)
  A_Tau <- 1 / pow(A_SD, 2 )
  A_SD ~ dgamma(1.01005, 0.1005012)
}
```